LINKAGE AND ASSOCIATION MAPPING OF SEED SIZE AND SHAPE IN LENTIL

A Thesis Submitted to the College of
Graduate Studies and Research in
Partial Fulfillment of the
Requirements for the Degree of
Masters of Science in the
Department of Plant Sciences
University of Saskatchewan
Saskatoon

By
Michael Fedoruk

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work, or in their absence, by the Head of the Department or Dean of the College in which my thesis was done. It is understood that any copying or publication or use of the thesis, in whole or in part, for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or in part should be addressed to:

Head of the Department of Plant Sciences
University of Saskatchewan
51 Campus Drive
Saskatoon, Saskatchewan
S7N 5A8

ABSTRACT

The seed size and shape of lentil are important traits because they determine the market class, cooking time, and can influence quality and yield of milled lentils. Understanding the genetic control of seed size and shape can help breeders develop varieties with improved seed size and shape characteristics such as seed diameter, seed thickness and seed plumpness. The objectives were to determine the heritability of seed size and shape and identify the genomic regions controlling these traits. This involved i) developing a linkage map for the LR-18 population (CDC Robin x 964a-46) using a recently developed single nucleotide polymorphism (SNP) assay; ii) analyzing the LR-18 population for seed size and shape QTLs; iii) analyzing an association mapping panel for seed size and shape QTLs. Phenotyping trials were grown at two different locations in Saskatchewan, Canada. The mapping population was grown in two different years while the association panel was only grown in one. Seed diameter and thickness were measured using sieves and this data were used to calculate seed plumpness. Days to flowering was also recorded to determine if it had any effect on seed size or shape. A linkage map consisting of 537 SNPs, 10 SSRs and 4 morphological markers on seven linkage groups was constructed and used for the QTL analysis. The heritability estimates were high for seed diameter and seed plumpness (0.92 and 0.94, respectively) while for seed thickness and days to flowering they were more moderate (0.60 and 0.45, respectively). QTL analysis revealed QTLs on five of the seven linkage groups. The association mapping study revealed similar heritability estimates of 0.97, 0.62, 0.94, and 0.62 for seed diameter, seed thickness, seed plumpness and flowering time, respectively. There were 31 different significant marker trait associations, however only 5 of those were significant for both locations. Four of those five markers did not map in the LR-18 linkage map so their genomic locations are still to be determined. Results showed that there are key regions in the genome that control seed size and shape and flowering time in lentil. These markers could be used for marker-assisted selection or for further candidate gene analysis.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Kirstin Bett for her mentorship, patience and the countless hours of discussion involving lentils, plant breeding and genomics. I would also like to thank my committee members Drs. A. Vandenberg, A. Beattie, R. McGee and B. Coulman for their advice and mentorship.

I am thankful to the Saskatchewan Pulse Growers for providing the funding needed for my project. I am also grateful to the Robert P. Knowles scholarship fund which provided me the financial support to complete this degree.

I would also like to extend my thanks to the many people in the Pulse Crop Breeding group at the CDC. This includes Brent Barlow, Thiago Prado, Robert Stonehouse, Marwan and countless others who provided me with technical assistance, guidance and friendship throughout my degree. My thanks also extend to the many graduate students in the Department of Plant Science whom I've been able to forge many close friendships with.

I'm also grateful to my family who provided me unconditional love and support.

TABLE OF CONTENTS

PERMISSION TO USE i

ABSTRACT ii

ACKNOWLEDGEMENTS iii

TABLE OF CONTENTS iv

LIST OF TABLES vii

LIST OF FIGURES viii

LIST OF APPENDICES x

LIST OF ABBREVIATIONS xi

Chapter 1 1

Introduction 1

Objectives and Hypothesis 2

Chapter 2 3

Literature Review 3

2.1 Lentil Genetics and Cytogenetics 3

2.2 Lentil Domestication 3

2.3 Lentil Production 4

2.4 Lentil Breeding In Canada 5

2.5 Breeding Lentil with Genebank Germplasm 7

2.6 Lentil Quality: Seed Size and Shape 8

2.7 Genetics of Seed Quality Traits in Lentil 9

2.8 Flowering Time and Seed Size 10

2.9 Lentil Genetic Linkage Mapping 11

2.10 Single Nucleotide Polymorphic (SNP) Markers for Lentil 12

2.11 Molecular Marker – QTL Associations 13

2.11.1 QTL Linkage Mapping 13

2.11.2 Association Mapping 15

2.11.3 Linkage Disequilibrium 16
2.11.4 Population Structure 19
2.11.5 Linkage and Association Mapping 21
Chapter 3 22
Construction of a Genetic Linkage Map 22
3.1 Introduction 22
3.2 Materials and Methods 22
3.2.1 Plant Material 22
3.2.2 DNA Extraction and SNP Genotyping 22
3.2.3 Linkage Map Construction 23
3.3 Results and Discussion 25
Chapter 4 31
Linkage Mapping of Seed Size and Shape QTLs 31
4.1 Introduction and Objectives 31
4.2 Materials and Methods 31
4.2.1 Plant Material 31
4.2.2 Phenotyping 32
4.2.3 Statistical Analysis 33
4.2.4 QTL Analysis 33
4.3 Results 33
4.3.1 Phenotypic Data 33
4.3.2 QTL Analysis 37
4.4 Discussion 41
Chapter 5 46
Association Mapping of Seed Size and Shape in Lentil 46
5.1 Introduction and Objectives 46
5.2 Materials and Methods 46
5.2.1 Plant Material 46
5.2.2 Phenotyping 47

5.2.3 Genotyping 47
5.2.4 Linkage Disequilibrium 48
5.2.5 Phylogenetic Tree Construction 48
5.2.6 Population Structure and Kinship Calculations 48
5.2.7 Association Analysis 49
5.3 Results 49
5.3.1 Phenotypic Data 49
5.3.2 SNP Genotyping 50
5.3.3 Linkage Disequilibrium 51
5.3.4 Population Structure 52
5.3.5 Association Analysis 56
5.4 Discussion 59
Chapter 6 65
General Discussion 65
6.1 Conclusions and Future Work 65
References 71
Appendices 84

LIST OF TABLES

Table 3.1. Markers used for the construction of the lentil linkage map. 26

Table 4.1. ANOVA results including F-values for seed diameter, thickness, plumpness and DTF for genotype and environmental effects. 34

Table 4.2. Pearson's correlation coefficients of seed traits and DTF, for each site-year 36

Table 4.3. Estimates of variance components and broad-sense heritability for seed size traits and DTF for RILs grown at two locations over two years (four site-years). 36

Table 4.4 QTLs identified for seed diameter, seed thickness, seed plumpness and DTF at four site years. 38

Table 5.1. Market classes of the CDC material used in this study. Many of the landraces do not fit the criteria for the Canadian market classes and as a result were not included in this table. 47

Table 5.2. F-values for seed diameter, seed thickness, seed plumpness and DTF. 49

Table 5.3. Variance components and broad-sense heritability of seed diameter, seed thickness, seed plumpness and DTF. 50

Table 5.4. Pearson's correlation coefficients for seed diameter, seed thickness, seed plumpness and DTF. 50

Table 5.5. SNPs used in the AM study that mapped in the LR-18 linkage map, and the number of SNPs with >10% allele frequencies. 51

Table 5.6. Significant marker associations with corrected p-values for seed diameter, seed thickness and seed plumpness estimated with the GLM model using SNP genotyping data for 140 diverse lentil lines. 58

Table 6.1. Significant markers in Chapter 5 and their positions on the LR-18 and LR-139 linkage maps. 69

LIST OF FIGURES

Figure 2.1. Comparison of QTL map resolution based on linkage mapping in a RIL population (left) and AM using a diverse collection (right) (reproduced with permission from Soto-Cerda and Cloutier, 2012). 16

Figure 2.2. Unlinked loci (θ=0.5) show higher levels of recombination leading to a more rapid LD decay over time. Highly linked loci (θ=0.0005) have a much slower rate of LD decay over time (reproduced with permission from Mackay and Powell, 2007). 17

Figure 2.3. LD decay over genetic distance in flax. The curved line represents a LOESS curve fit for the scatterplot (reproduced with permission from Soto-Cerda and Cloutier 2012). 18

Figure 3.1. Example of SNP analysis using GenomeStudio ver 2010.1. The coloured dots correspond to each individual in the population. Monomorphic markers (A) appear clustered in the same region (blue), while polymorphic markers (B) show 1:1 segregation for each allele (blue and red). Heterozygous markers are located in between the two groups (pink). Markers not belonging to a distinct group are not coloured (black). 23

Figure 3.2. Dotplot showing collinearity between lentil and *Medicago*. The marked circles represent translocations. Linkage groups for lentil were selected based on their match with the chromosomes in *Medicago*. (reproduced with permission from Sharpe et al. 2013)..... 25

Figure 3.3 Linkage map of the LR-18 population. Seven linkage groups are shown which correspond to the seven chromosomes of lentil. 28

Figure 4.1. Differences in seed size and shape among the parents and two randomly chosen RILs from the LR-18 population. 32

Figure 4.2. Box and whisker plots of the distribution of seed diameter, thickness and plumpness, and DTF in the LR-18 (CDC Robin x 964a-46) RIL population grown at Preston and SPG in 2009 and 2011. The mean value of the parents are labeled with an arrow. 35

Figure 4.3. Genetic linkage map of lentil RIL population LR-18 with QTLs for seed diameter, seed thickness, seed plumpness and DTF. All linkage groups are shown with QTLs marked next to the loci they are associated with. The thin lines represent the regions that were significant using interval mapping, while the boxes represent

the regions significant under composite interval mapping. QTL boxes marked with an asterisk (*) represent QTLs that were significant for all site-years..... 40

Figure 5.1. Example of differences in seed size and shape among the material grown in the association panel. 47

Figure 5.2. Linkage disequilibrium (r^2) plotted against genetic distance (cM) for pair-wise comparisons of markers located throughout the genome. The red lines indicate the second-degree LOESS that was fit for each plot and help represent the rate of LD decay. The green dashed line represents the fixed r^2 value of 0.1. Any value above 0.1 is considered in LD..... 52

Figure 5.3. Subpopulations and admixtures of 140 lentil lines genotyped and sorted into populations based on STRUCTURE analysis. Each bar represents the individual while the color represents the subpopulation and admixture of each individual..... 54

Figure 5.4. Contribution of each ancestral population to the STRUCTURE groups. Group 1 consists of elite breeding lines but does carry a significant amount of landraces. Groups 2 and 4 mainly consist of landraces while group 3 is predominantly elite breeding lines..... 54

Figure 5.5. Distribution of seed diameter for each of the population groups identified in STRUCTURE. 55

Figure 5.6. UPGMA dendogram of the lentil diversity panel constructed using NEI's (1972) standard genetic distance measurement method. Regions surrounded by different colours correspond to the different sub-groups constructed using STRUCTURE. Un-highlighted clusters were admixtures indicating hybrids between groups..... 56

Figure 5.7. Cumulative p-value distributions for the three different association models used for each location (SPG and Preston). A model without any population structure or kinship control (none) is compared to a generalized linear model (GLM) with population control and a mixed linear model with population and kinship control (MLM). 59

LIST OF APPENDICES

Appendix 1. Association panel number, accession name, country of origin and their STRUCTURE sub-group assignment. 84

Appendix 2. STRUCTURE sub-groups and each line's assigned groupings. Each value within the sub-groups shows the level of admixture that each accession has for the sub-groups. 88

LIST OF ABBREVIATIONS

AM	Association Mapping
CDC	Crop Development Center
CIM	Composite Interval Mapping
DTF	Days to 50% Flower
EST	Expressed Sequence Tag
GLM	Generalized Linear Model
GWAS	Genome Wide Association Mapping
ICARDA	International Center for Agriculture Research in the Dry Areas
IM	Interval Mapping
LD	Linkage Disequilibrium
LOD	Logarithm of Odds
MIM	Multiple Interval Mapping
MLM	Mixed Linear Model
PI	Plant Introduction
QTL	Quantitative Trait Loci
RIL	Recombinant Inbred Line
SNP	Single Nucleotide Polymorphism
USDA	United States Department of Agriculture

Chapter 1

Introduction

Lentil (*Lens culinaris* ssp. *culinaris* Medik.) is a crop that is consumed for its high levels of protein and micronutrients including iron, zinc and β-carotene (Erskine and Sarker 2004). Lentil has become an important crop for western Canadian growers with acreage steadily increasing. Canadian lentil acres have gone from 750,000 acres in 1996 to more than 2 million acres in 2011 (FAOSTAT 2010). Maintaining the quality of lentils for the end users is an important objective for the industry. The size and shape of the lentil is considered an important parameter in reaching optimum quality. This is because the size and shape of lentils can influence the cooking time and dehulling efficiency and can be valued specifically to market preferences (Erskine et al. 1991; Wang et al. 2008)

Developing new and improved seed sizes and shapes is an important objective for the lentil breeding program at the Crop Development Center (CDC) located at the University of Saskatchewan. Currently, at the CDC, selection for seed size and shape is done through phenotypic evaluation. Molecular markers that are linked to the seed size and shape traits could also be used to select for those traits. This could increase the rate at which new cultivars with new seed sizes and shapes are released. However, in order for molecular markers to become implemented, experimental populations need to be developed and evaluated for seed size and shape to draw statistical associations between those traits and molecular markers.

A lentil recombinant inbred line (RIL) population segregating for seed size and shape has been developed. An association mapping (AM) panel, which contains cultivars, breeding lines and landraces that also differ in seed size and shape, has also recently been developed. Evaluating both the RIL population and association panel with single nucleotide polymorphism (SNP) markers can allow for the association of a marker, or specifically an allele, to a phenotype. For seed size and

shape, this will investigate the genetic control of these traits along with the development of molecular markers that can be used by the breeding program.

Objectives and Hypothesis

The objective of this project is to optimize the use of SNP markers in lentil and enhance the understanding of the genetic control of seed size and shape in lentil with results leading to the development of molecular markers that can be adopted by breeding programs. There are three different aspects to this study: 1) genotyping a RIL population and association mapping panel with SNP markers, 2) phenotyping both populations for seed size and shape, and 3) identifying significant marker-trait associations via quantitative trait loci (QTL) mapping (in the RIL population) and association mapping.

The hypothesis for this research is that genomic regions controlling seed size and shape of lentil can be identified through linkage analysis and that association mapping will confirm these QTLs and also yield new associations to alleles that control unique seed sizes and shapes.

Chapter 2

Literature Review

2.1 Lentil Genetics and Cytogenetics

Lentil (*Lens culinaris* ssp. *culinaris*) is a diploid self-pollinating crop with seven chromosomes (2n=14). It has genome size of 4063 Mbp/1C (Arumuganathan and Earle 1991). There are five other species in this genus: *L. odemensis*, *L. ervoides*, *L. nigricans*, *L. tomentosus, L. lamottei* (van Oss et al. 1997).

2.2 Lentil Domestication

Large chromosomal variations exist amongst all the species within *Lens*. However, Ladizinsky (1979) found that there were fewer chromosomal interchanges between *L. culinaris* ssp. *culinaris* and *L. culinaris* ssp. *orientalis* versus *L. culinaris* ssp. *culinaris* and the other *Lens* species. L. *culinaris* ssp. *culinaris* and L. *culinaris* ssp. *orientalis* also share similar morphology and molecular marker genotypes (Hancock 2004). As a result, *L. culinaris* ssp. *orientalis* is considered to be the most likely progenitor of *L. culinaris* ssp. *culinaris* (Sonnante et al. 2009; van Oss et al. 1999).

Barulina (1930) first suggested that lentil was domesticated in the Hindu-Kush region of central Asia. However, subsequent archaeological studies have shown that lentils were more likely domesticated in the Fertile Crescent, of modern day southern Turkey and Syria, 10,000 years ago. Lentil then spread from this region west along the European Mediterranean coast, east towards the Indian sub-continent and south into Egypt's Nile valley (Erskine and Sarker 2004).

The domestication of lentil involved modifications to traits like pod dehiscence and seed dormancy, which allowed the seeds to be collected more easily and meant that farmers were able to keep their seeds to the next growing season (Sonnante et al. 2009). It is also believed that the domestication of lentil led to changes in seed size. Wild lentils have much smaller seeds, which would have been difficult to collect; therefore increases in seed size would have made it easier for

humans to harvest the seeds. Almost all the domestication traits of lentil, such as pod dehiscence, seed dormancy, and growth habit, involved single gene inheritance, which would allow mutations in those genes to be selected for and retained more easily. Seed size, on the other hand, has a more complex, quantitative mode of inheritance. As the cultivation of lentil continued to spread, the species diverged in to different sub-groups. These subgroups differed mainly in their seed size and are known as microsperma and macrosperma. The microsperma types have a seed diameter of 2 to 6 mm, have red and yellow cotyledons, and pigmented flowers. The macrosperma types, meanwhile, have a seed diameter of 6 to 9 mm with a yellow cotyledon and no pigmentation in the flowers (Sandhu and Singh 2007). Barulina (1930) was the first to characterize lentils based on these attributes. She also highlighted that their geographic origins differ with the microsperma types being centered in southeast Asia and the macrosperma types more common in western Asia and Europe.

2.3 Lentil Production

Today, lentil is grown in temperate to sub-tropical regions throughout the world. It is still grown in the traditional regions of southern Europe, the Middle East, northeastern Africa and the Indian sub-continent. However, lentil production has now spread into growing regions of South and North America and Australia.

Canada is the number one producer and exporter of lentils in the world. In 2009 Canada exported 1.2 Mt of lentils. The province of Saskatchewan accounts for 99% of Canadian production (Saskatchewan Pulse Growers 2012). The United States is the second largest exporter with 0.18 Mt exported in 2009, followed by Turkey with 0.13Mt exported. (FAOSTAT 2010). India is the largest importer of lentils with 0.29 Mt imported in 2009. Bangladesh and Turkey are the second and third largest importers with total imports reaching 0.17 and 0.14 Mt, respectively.

In Canada, there are many different market classes that determine lentil value. The largest are the green and red lentil market classes. For green lentil, there are small, medium and large sub-classes. Green lentils, usually large in size, have seed weights >6g/100 seeds, green seed coats, yellow cotyledons and are normally

cooked and consumed whole (Erskine 1996). Red lentil, which is much smaller, has medium, small and extra small market classes. Red lentils typically are <3.5g/100 seeds, have a brown to grey seed coat and red cotyledons. Red lentils are traditionally dehulled and spilt. Dishes containing dehulled spilt red lentils are traditionally named "dhal". There are also specialty or niche market classes like the French green, Beluga, Pardina, zero tannin and green cotyledon market classes. (Saskatchewan Pulse Growers 2012).

2.4 Lentil Breeding In Canada

Lentil breeding started in Canada with the appointment of a breeder at the Crop Development Center (CDC) at the University of Saskatchewan in 1972 (Morrall 1997). To this day, the lentil breeding program at the CDC remains the only one in Canada. The first cultivar that was registered was Laird, in 1978. Laird was selected from the United States Department of Agriculture Plant Introduction (PI) line 343028, which originally came from Russia (Slinkard and Bhatty 1979). The second lentil cultivar, Eston, was released in 1980. Eston was selected from the accession PI 179307, which originally came from Turkey (Slinkard 1981). The cultivar Rose, was the first red cotyledon variety to be released and also the first Canadian cultivar to be released that originated from a cross (Eston x Redchief) (Slinkard and Vandenberg 1995). CDC Gold, the first zero tannin variety, was released in 1993. Other notable varieties include CDC Imperial and CDC Impact, which were the first imidazolinone (IMI) tolerant lentil cultivars, released in 2006.

Breeding methods for lentil at first were pure-line selections from landrace accessions. The F_2 derived family method , a modified form of bulk selection, involves bulking F_3 plants derived from selected single F_2 plants followed by yield testing of selections in the F_4 –F_8 generations. This method was used for the early cultivars that were developed through crossing (Slinkard and Vandenberg 1995) and is still the preferred breeding method. Backcrossing was first used for the development of IMI tolerant cultivars. CDC Imperial and CDC Impact were the result of a backcross with the IMI tolerant breeding line RH44 to the cultivars CDC Robin and CDC Blaze, respectively (Chant 2004). Throughout the breeding program, most

of the selections, like yield and seed coat colour, are made through phenotypic evaluation. When IMI tolerant varieties were introduced, IMI herbicides were applied to the segregating material to select for tolerance.

For many other crops, molecular markers have been used to select for traits that are quantitatively inherited, have low heritability or are difficult to select for by phenotype alone. Examples are: selecting for common bacteria blight (CBB) resistance in common bean, and for quality traits like low cadmium in durum. (O'Boyle et al. 2007; Wiebe et al. 2010) Molecular markers have been developed in lentil that are linked to disease resistance loci, frost tolerance, and flowering related traits (Ford et al. 1999; Tar'an et al. 2003; Kahraman et al. 2004; Tullu et al. 2008). The low efficacy and the difficulties of using some of those markers have prevented the breeding program at the CDC from adopting marker-assisted selection. Recently, there has been increased funding available for research in lentil genomics that should result in numerous functional molecular markers in the near-term. Next-generation technologies have also decreased the cost and increased the efficiency of marker-assisted selection. As a result, the lentil breeding program at the CDC is in the process of developing a marker-assisted breeding strategy using the latest genomics technologies.

The use of molecular markers will complement the F_2–derived family breeding method currently being used. Multiple parents are used in many of the crosses that are made. Gamete selection, using molecular markers, would allow for the selection of preferred F_1 progeny resulting in populations that are enriched for desirable alleles (Singh 1994). Molecular markers that are tagged to specific traits could also facilitate more efficient breeding with un-adapted or wild germplasm. Functional markers would result in less linkage drag when backcrossing with exotic breeding germplasm. However, for this to occur, genetic variation for the traits of interest needs to be measured and associated with molecular markers. QTL mapping in bi-parental populations and association mapping are two methods which can be employed to achieve this.

2.5 Breeding Lentil with Genebank Germplasm

In total, over 43, 000 accessions of *Lens* have been collected and deposited in genebanks around the world. The largest is the International Center for Agriculture Research in the Dry Areas (ICARDA) genebank, which holds over 11,000 landraces (*L. culinaris* ssp. *culinaris*) and wild species of *Lens* (Global Crop Diversity Trust 2012). The other major collection is located at the USDA-ARS germplasm repository in Pullman, Washington. This collection has around 10,800 accessions, of which 8,860 are landraces of L. *culinaris* ssp. *culinaris.* Other large collections also exist in Australia and Iran (Global Crop Diversity Trust 2012). At the University of Saskatchewan a mini core collection was assembled, consisting of landraces and wild *Lens* accessions from the ICARDA genebank. This core collection contains material that is more adapted to western Canadian growing conditions (A. Tullu, pers. comm.).

Broad phenotypic variation exists within these collections, especially amongst the wild *Lens* accessions. Resistance to diseases such as ascochyta blight, anthracnose and vascular wilt has been identified in some *Lens* accessions (Tullu et al. 2006 and 2010; Bayaa et al. 1995). Seed composition traits that influence quality have also been studied in these collections (Tahir et al. 2012). Plant breeders could use this variability to develop improved cultivars by crossing with elite material. However, integration of variation from different *Lens* species into the cultivated lentil genetic background has proven to be difficult. Interspecifc hybrids can be difficult to produce and, when successfully produced, can be sterile (Ladizinsky et al. 1985). Breeders are also aware of linkage drag, where additional genetic material is introduced along with the desired gene of interest from the wild plant source, causing a negative impact on plant performance. A more reasonable approach to increasing the variability within breeding programs may be introgressing variation from lentil landraces or other material from independent breeding programs. Landraces of lentil have shown to be resilient to drought based on the wide range of growing conditions under which they can survive (Muehlbauer et al. 1995). Lazaro et al. (2001) observed that Spanish lentil landraces exhibit variability for traits including plant height, flower duration, days to maturity and seed weight.

Significant variation in quality traits such as sucrose concentration and raffinose family oligosaccharides (RFOs) were also observed within a mini-core collection (Tahir et al. 2012). This suggests that broad phenotypic variation exists within lentil landraces that it could be an important resource in further breeding efforts.

2.6 Lentil Quality: Seed Size and Shape

Improved quality is an important objective for lentil breeders. Quality can be determined by characteristics like size, shape, colour, taste, and cooking time. Seed size is an important quality trait in many crops. For wheat, increases in seed size can increase the yield obtained in the milling process (Breseghello and Sorrells 2006). In soybean (*Glycine max* L.) seed size largely determines the end-uses of the seed. Small and large sizes are consumed as food, while medium sizes are crushed for their oil and meal (Shanin et al. 2006). Seed size, which is measured as the seed diameter, is also an important trait in a crop like lentil. First, seed size determines the amount of time it takes to cook the lentils. Hamdi et al. (1991) observed a strong positive correlation, of r=0.96, between seed size and cooking time. Also, size can affect the outcome of the seed when it is handled, processed and spilt. Ford et al. (2007) noted that development of rounder shaped lentil cultivars, versus the usual thin, sharp-edged types, could reduce the amount of damage that occurs during handling. Past studies have also highlighted that rounded or plumper lentils exhibit greater dehulling efficiency versus thinner, less plump samples thus increasing the value of the crop (Erskine 1991; Wang et al. 2008; Shanin et al. 2012). If plant breeders can capitalize on certain market preference for specific seed sizes, more value could be added to the crop.

Determining seed size in lentil has historically relied on measuring the 100 or 1000 seed weight (Erskine et al. 1985; Abbo et al. 1991; Tahir et al. 1995; Tullu et al. 2001). However, this method cannot distinguish different seed shape parameters such as seed thickness or seed plumpness. Several different methods have been used to measure the size and shape of legume seeds. Xu et al. (2011) randomly selected soybean seeds and measured various seed shape parameters using a caliper. Computer-assisted two-dimensional image analysis has been used to

measure the distribution of lentil seed diameter (Shanin and Symons 2001). Shanin et al. (2006) used cameras to capture the 3-dimensional image of lentil seeds to determine the plumpness. However, these methods are very laborious when working with large populations. Hossain et al. (2010) noted that the traditional method of seed sizing using graded sieves can be just as effective in determining seed size and shape.

Previous genetic studies in lentil revealed that there is large variation for seed weight (Tullu et al. 2001; Abbo et al. 1991). Considerable variation was also observed for seed diameter ranging from 3 to 9 mm by these researchers. No previous studies have evaluated seed thickness and seed plumpness in lentil. For other crops like soybean and wheat, it has been reported that there is no association or linkage between seed size and shape (Cober et al. 1997; Gegas et al. 2010).

2.7 Genetics of Seed Quality Traits in Lentil

Seed quality is an important characteristic for meeting market demands. The seed coat colour and pattern, cotyledon colour along with the size and shape of the seed are the traits that can determine the value of a lentil variety. Nearly all seed quality traits, except seed size and shape, are qualitatively inherited. Vandenberg and Slinkard (1990) first determined the inheritance of seed coat colour and seed coat pattern in lentil. For seed coat colour, it was determined that there were two independent loci controlling whether the seed coat would be grey (*Ggc*, *tgc*) or tan (*ggc,Tgc*). When both dominant alleles are present (*Ggc*, *Tgc*) a brown seed coat is produced. Double homozygous recessive individuals (*ggc*, *tgc*) have a green seed coat. Seed coat pattern was determined to have five different alleles, marbled1 (Scp^{m1}), marbled2 (Scp^{m2}), spotted (Scp^{s}), dotted (Scp^{d}) and no pattern (*scp*) all at one locus. The inheritance of cotyledon colour in lentil was described by Slinkard (1978). Crosses between red and yellow cotyledons indicated single gene inheritance, with a 3:1 ratio of red to yellow cotyledons. The designated gene symbols were *yc* for the yellow cotyledon, *Yc* for red cotyledon and *i-yc* for green cotyledon. Eujayl et al. (1998) mapped the *Scp* locus, within close linkage to another qualitative trait, flower colour (*W*). Duran et al. (2004) mapped the cotyledon colour

and seed coat colour and pattern loci to three different regions of the genome. Fratini et al. (2007) mapped the *Scp* and *Yc* loci to two separate linkage groups.

The other important seed quality traits, seed size and shape, are quantitatively inherited. There are multiple loci that can control seed size and shape and there are many different environmental and physiological factors that can affect seed size in crop plants. Quantitative trait loci (QTL) for seed weight in lentil were located by Abbo et al. (1991), but simply by measuring seed weight it is difficult to differentiate whether the seed is plumper or just has a larger seed diameter. Fratini et al. (2007) mapped QTLs for seed diameter, seed weight and flowering time, among other traits. In this study seed weight and seed diameter were significantly correlated, but the QTL were not co-located. Also, because their experimental population was derived from cross between two different sub-species (*L. culinaris* ssp. *culinaris* x *L. culinaris.* ssp. *orientalis*), this would make using those markers difficult for MAS within domesticated material.

2.8 Flowering Time and Seed Size

Recent studies of seed size QTLs in other crops have also uncovered loci controlling more than one related trait. In lentil, a flowering time locus has been shown to be linked with the seed coat pattern locus (Sarker et al. 1999). Flowering, or more specifically, pre-anthesis and post-anthesis periods have been shown to also have an influence on seed size (Gupta et al. 2006). For example, pre-anthesis changes in vegetative organs can affect the amount of assimilates that are partitioned to the seed while they are developing. Similarly, post-anthesis processes can affect the time for maturation or grain filling which could change the seed size. Loci controlling flowering time or other flower morphology traits have also been associated with seed mass or seed size loci in model legume crops (Ohto et al. 2005; He et al. 2010; Wang et al. 2012). In chickpea, Hovav et al. (2003) studied how a major flowering time gene, PPD, affected the seed weight. They found that earlier flowering resulted in reduced seed weight, which could lead to lower yields and quality. This would affect how cultivars are selected, because it could be difficult to select for early flowering without affecting seed size. Therefore it is important to

note that when selecting markers linked to QTLs for seed size, some other traits can be influential or even mask the QTLs of interest.

2.9 Lentil Genetic Linkage Mapping

Linkage mapping is the process of assembling marker genotypes (morphological and molecular), determining the distances between them through measuring recombination in a population, and then placing them into linkage groups (Jones et al. 1997). Linkage maps are useful for locating quantitative traits and understanding the genetic make-up of a crop. Many of the genetic maps in lentil were developed from interspecific populations between *L. culinaris* ssp. *culinaris* and other *Lens* species (Zamir and Ladizinsky 1984; Tadmor et al. 1987; Havey and Muehlbauer 1989; Tahir et al. 1993; Eujayl et al. 1997; Duran et al. 2004). These populations were used to increase the probability of detecting polymorphisms between the parents (Ford et al. 2007). However, when working with interspecific populations, segregation distortion, caused by favouring one parental allele over the other is common, as is inaccurate estimation of map size. Additionally, the markers that are polymorphic within interspecific populations are often not polymorphic in the cultivated species genetic background, limiting their utility in breeding programs, and other studies using only cultivated germplasm (Ford et al. 2007). However, recent developments in marker technologies have resulted in an increasing ability to detect polymorphism between potential parents within the cultivated species *L. culinaris* ssp. *culinaris*. Rubeena et al. (2003) were the first to develop an intraspecific mapping population in lentil. Subsequently, there have been a number of intraspecific populations developed for lentil with the purpose of QTL mapping (Kahraman et al. 2004; Phan et al. 2007; Tullu et al. 2008).

Morphological, allozyme and isozyme markers were initially used for mapping studies, but now DNA-based markers are primarily used. Amplified fragment length polymorphisms (AFLPs), restriction fragment length polymorphism (RFLPs), and random amplified polymorphic DNA (RAPD) markers have all been used in previous mapping studies (Eujayl et al. 1997; Rubeena et al. 2003; Duran et al. 2004; Fratini et al. 2007). However, these markers are either anonymous, which

means that they are in a non-coding genomic region, are not associated with a change in amino-acid sequence, or are not amenable to high-throughput MAS (Batley and Edwards 2007). Phan et al. (2007) used intron targeted amplified polymorphic (ITAP) markers to develop the first gene-based lentil linkage map. However, only 79 of these markers were used in the map, leaving large gaps in the genome. Single sequence repeats (SSRs) are single locus, multi allelic markers that have been used in many mapping projects and have been the preferred marker of choice in marker-assisted selection in other species (Ford et al. 2009). SSRs have been used in both mapping and diversity studies in lentil (Fratini et al. 2007; Tullu et al. 2008; Liu et al. 2008; Babayeva et al. 2009; Hamwieh et al. 2009). More recently, Gupta et al. (2012) developed a linkage map based on 196 SSRs markers, 15 of which were derived from EST sequences. This is a very limited number of markers for lentil, especially considering the size of the genome.

Vail (2010) noted that there is limited routine application and reproducibility of many of these markers in lentil breeding and genetic studies. This may be because some of these markers are specific to certain genetic backgrounds, and/or are difficult and expensive to screen (Ford et al. 2009). In order to develop high-density linkage maps in lentil, more robust and abundant markers need to be developed.

2.10 Single Nucleotide Polymorphic (SNP) Markers for Lentil

Single nucleotide polymorphic (SNP) markers are the most abundant type of polymorphic marker that can be found within the genome of a species. Recent developments in sequencing technology have allowed for the discovery of large numbers of SNPs in many different crop species. There are different routes that can be taken for SNP discovery, all of which require reliance on sequence information. Re-sequencing PCR amplicons, electronic SNP discovery in shotgun genomic libraries or discovery in expressed sequence tag (EST) libraries are all methods for SNP discovery (Rafalski 2002). SNPs have already been used to increase the resolution of numerous genetic maps for various crops (Gupta et al. 2008).

The recent development in high-throughput highly parallel multiplex assays, or chips, is allowing for the genotyping of many individuals with hundreds to thousands of SNP markers at one time. These types of platforms have already been developed for numerous crops such as barley, corn, and apple (Close et al. 2009; Yan et al. 2009; Chagne et al. 2012). Recently, a highly parallel allele-specific Illumina Golden Gate 1536-SNP assay has been developed for lentil (Sharpe et al. 2013). This array was constructed using SNPs discovered in expressed sequence tag (EST) sequences from nine *L. culinaris* ssp. *culinaris* and two *L. ervoides* accessions. This array is based on Illumina's BeachChip™ technology where allele specific oligonucleotides that are fluorescently labeled bind to specific SNP alleles. The reagents are run through a PCR reaction with the template DNA and then hybridized to the bead chip (Fan et al. 2006). The products of the reaction are then read on an Illumina HiScan scanner (Illumina, San Diego CA). This particular lentil array has 1,536 bi-allelic SNPs that can be screened across mapping populations, association panels or any collection of lines.

2.11 Molecular Marker – QTL Associations

Molecular markers can be a valuable tool for plant breeders. If a molecular marker is inherited together with, or is significantly associated with, a particular trait, breeders could use them to predict the phenotypic value of an individual. This allows for breeders to select traits that have low heritability and also increase the efficiency of their programs by having the desired individuals selected before they enter the field. QTL mapping and association mapping are two methods that are widely used among plant geneticists to associate traits with molecular markers.

2.11.1 QTL Linkage Mapping

QTL mapping is a strategy that detects associations between a quantitatively inherited phenotype and markers. QTL linkage mapping relies on genetic markers, which have been placed in a linkage map, to associate any genetic variation, based on familial relationships, with phenotypic variation that has been measured in multiple environments. If a particular marker can be associated with a phenotype, it can be potentially used in MAS. Different types of bi-parental populations, such as

F_2, doubled haploids (DH), recombinant inbred lines (RILs), backcrosses and near isogenic line (NILs) are commonly used for linkage mapping. Using maximum likelihood or regression analyses, models such as interval mapping (IM), composite interval mapping (CIM) and multiple interval mapping (MIM) have been used to determine associations between the markers and traits of interest. CIM can account for multiple QTLs, while MIM can account for multiple QTLs and their epistatic effects (Semagn et al. 2010).

QTL studies using linkage mapping are abundant in nearly all crop species, including lentil. Multiple QTLs in lentil have been identified and mapped, using both inter-and intra-specific maps. Of the QTL mapped so far, many have been for agronomic traits such as plant height, days to flowering, winter hardiness, pod dehiscence, grow habit and yield (Tullu et al. 2008; Kahraman et al. 2004; Fratini et al. 2007). QTLs for resistance to diseases like ascochyta blight, anthracnose and stemphylium blight have also been mapped (Ford et al. 1999; Rubeena et al. 2006; Tullu et al. 2006; Saha et al. 2010). QTLs for seed quality traits, such as seed size and shape, are limited. QTLs for seed weight in lentil have been located by Abbo et al. (1991) and Fratini et al. (2007) mapped QTLs for seed diameter and seed weight. Even with the number of QTLs that have already been mapped, very few markers are currently being used for MAS in lentil breeding.

Genetic linkage mapping has a number of shortcomings. The experimental populations used in linkage mapping lack the number of recombination events normally seen in natural populations, which lead to QTLs with poor map resolution. Many QTLs can span distances up to, or even greater than 10cM (Holland 2007). Numerous genes could exist within those regions, leading to speculation on the exact location of the QTL. In addition, QTLs that span large distances can become disassociated from identified markers, in other populations or breeding material, due to historical recombination that may exist within the germplasm. The experimental populations used for linkage mapping are generally the result of a cross between two individuals. Only the alleles present and polymorphic between the parents of that cross can be sampled in that population. The genetic variation that exists between two individuals does not constitute the entire genetic variation

for a given crop species, therefore the genetic variation is under-represented. This could lead to QTLs that may only exist within the specific experimental populations, and not among the rest of the germplasm for that crop species. MAS would thus not be effective when other lines are evaluated. In order to overcome the limitations of QTL discovery via linkage mapping, many populations could be developed to assemble all the alleles for that crop species. However, each population would need to be created, genotyped and phenotyped which would become expensive and laborious.

2.11.2 Association Mapping

Association mapping (AM) is a method that can address the shortcomings of linkage mapping. This method allows the use of many diverse individuals, which increases the number of alleles examined and samples multiple historical recombination events. As a result, properly chosen AM panels have a greater frequency of alleles that encompass the genetic variation of the crop species. This can reduce the time, along with the costs, to identify markers linked to quantitative traits. Association mapping capitalizes on the historical levels of recombination accumulated in natural populations, landraces, breeding material and varieties, which results in higher QTL resolution than linkage mapping (Figure 2.1). These advantages have made AM a valuable method in marker-trait associations.

There are two different types of association mapping reported in the literature: genome-wide association studies (GWAS) and candidate gene association mapping. The GWAS method scans the entire genome to determine if any association between markers and phenotypes exists. This method requires that there are enough markers to cover the genome based on the expected rate of linkage disequilibrium (LD) decay. The other method, candidate gene AM, requires prior knowledge of candidate genes that could be associated with a phenotype. This knowledge may have been gained through QTL linkage mapping, GWAS or from work in related species. Instead of a whole genome scan, only markers within those candidate genes are analyzed for associations.

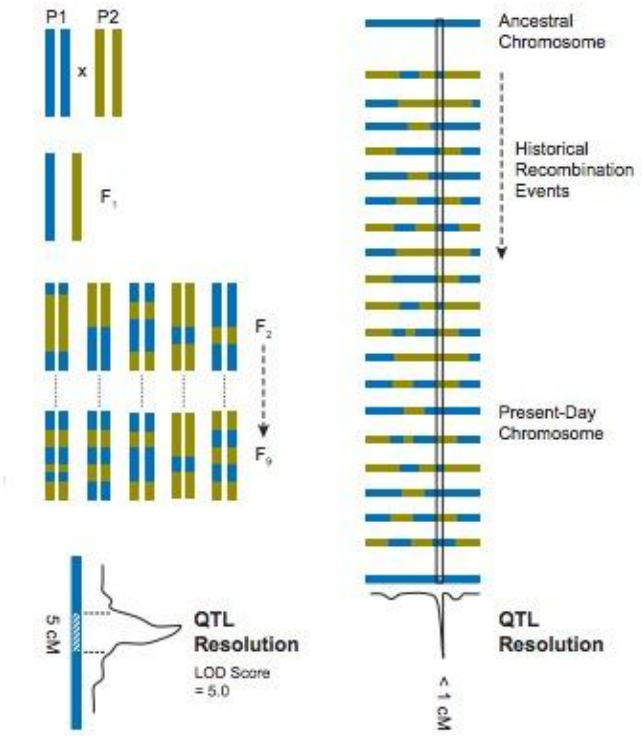

Figure 2.1. Comparison of QTL map resolution based on linkage mapping in a RIL population (left) and AM using a diverse collection (right) (reproduced with permission from Soto-Cerda and Cloutier, 2012).

2.11.3 Linkage Disequilibrium

Linkage disequilibrium (LD) is the non-random association of alleles at different loci (Oraguzie et al. 2007). AM, which sometimes is referred to as LD mapping, statistically associates markers that are in LD with the genetic variants of a phenotype. The level of LD determines the resolution of association mapping studies. A high LD means a lower resolution, while lower LD means greater resolution. LD can be affected by many factors including the amount of inbreeding, population size, genetic isolation between lineages, population subdivision, recombination rates, population admixtures, mutations, and whether individuals have been under natural or artificial selection since diverging (Gupta et al. 2005; Mackay and Powell 2007). All these factors affect how LD decays over time, with loci displaying lower levels of recombination (e.g. $\theta = 0.0005$) showing lower LD decay, while individuals displaying higher recombination levels (e.g. $\theta = 0.5$) having higher LD decay (Figure 2.2). Natural and wild populations usually exhibit lower levels of

LD versus cultivated or elite breeding germplasm due to their greater levels of recombination over time. Natural and wild populations have gone through little artificial selection pressure. They also tend to have more diverse alleles per locus because these populations have not gone through any of the genetic bottlenecks that are observed due to domestication and selection. This would result in more polymorphic markers located closer to the gene responsible for the phenotypic variability, hence increasing the mapping resolution and the stability of the marker.

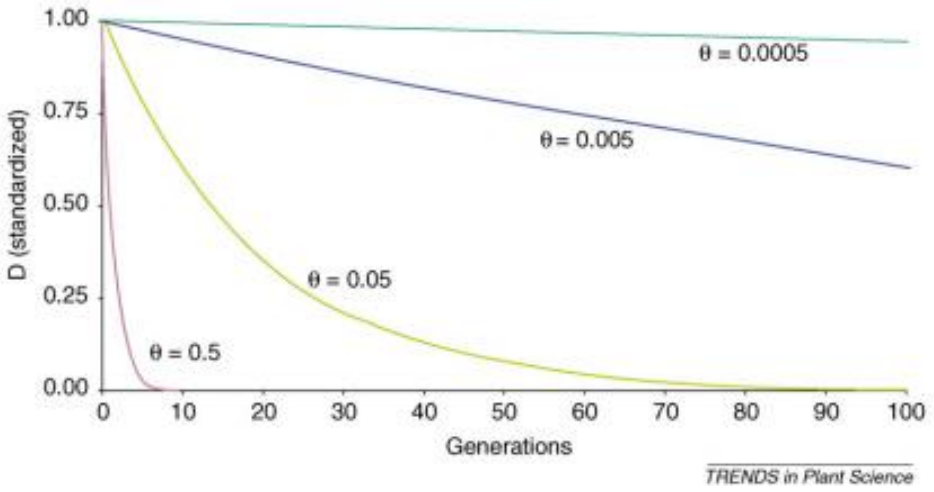

Figure 2.2. Unlinked loci (θ=0.5) show higher levels of recombination leading to a more rapid LD decay over time. Highly linked loci (θ=0.0005) have a much slower rate of LD decay over time (reproduced with permission from Mackay and Powell, 2007).

LD is measured as the difference between observed and expected gamete haplotype frequencies under linkage disequilibrium (Soto-Cerda and Cloutier 2012). LD can be measured using either the D' or r^2 measurements. In most studies, r^2 is preferred to measure LD through pair-wise measurements between markers. This is because D' can be inflated by small sample size and low allele frequencies, while r^2 shows less bias (Soto-Cerda and Cloutier 2012). LD decay can be visualized by measuring LD over genetic or physical distance using LD scatterplots (Figure 2.3). This method can help researchers understand the level of LD over chromosomes or even full genomes. By understanding the rate of LD decay over distance, the number of markers that would be needed for GWAS could be determined. LD heat maps are

also used to visualize LD over a single gene, or entire chromosome. This method can be useful in determining which regions, or loci, of a chromosome are under greater levels of LD.

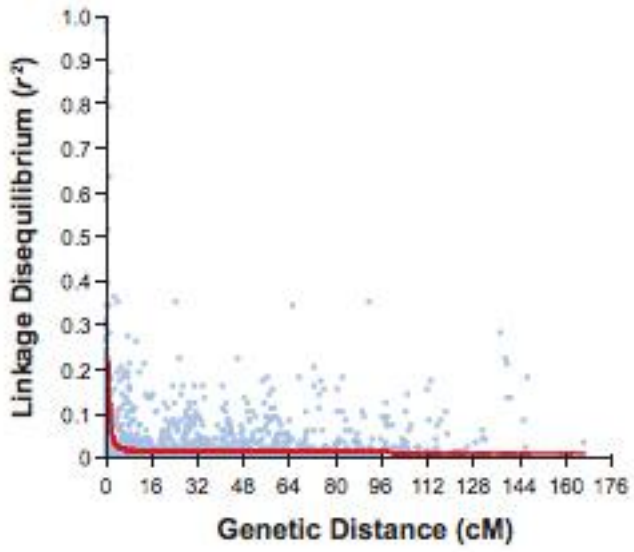

Figure 2.3. LD decay over genetic distance in flax. The curved line represents a LOESS curve fit for the scatterplot (reproduced with permission from Soto-Cerda and Cloutier 2012).

It is important to measure LD prior to AM in order to determine how many markers are needed to saturate the genome for GWAS. LD has shown to be variable among species. For example, corn, which has low LD because it is an outcrossing species, exhibits LD decay within a few hundred base pairs (Tenaillon et al. 2001). This makes genome-wide association mapping difficult because many markers are needed to account for all the small LD blocks, resulting in increased genotyping cost. As a result the candidate gene AM approach is predominately used in corn and other species where LD has been shown to decay fast. In other crop species, such as barley and soybean, which are self-pollinated, the extent of LD is greater making GWAS more feasible. For example, Hamblin et al. (2010) estimated genome wide LD decay to be 20-30 cM among elite barley cultivars.

To date there are no reported association mapping studies in lentil primarily due to the lack of genomic resources available for lentil. However, this changed when Sharpe et al. (2013) developed a 1536-SNP Illumina GoldenGate array

(Lc1536). It is anticipated that a self-pollinated crop like lentil will exhibit levels of LD similar to barley and soybean, making the Lc1536 array appropriate for GWAS. However, in the model species *Medicago truncatula*, to which lentil is often compared, LD has been noted to decay within 3kbp (Branca et al. 2011). Therefore, understanding the level of LD in lentil will be an important first step for association mapping in lentil.

2.11.4 Population Structure

Population structure is a constraint that can create false associations in association mapping studies. Population structure is formed by non-random mating within a species, which can cause changes in allele frequencies (Ersoz et al. 2007). The non-random mating may be due to events such as genetic drift and domestication bottlenecks. This can inflate the presence of certain marker alleles resulting in overrepresentation in a population, which in turn cause them to be falsely associated with a phenotype (Pritchard et al. 2000). There are different statistical approaches for controlling population structure in association mapping. In population based studies, two approaches are used: genomic control (GC) and structured association (SA) (Yu and Buckler 2006). Genomic control calculates the non-independence of loci, which corrects for any population structure (Ersoz et al. 2007). The significance tests, or P-values, are then adjusted to account for the population structure. However, as Mackay and Powell (2007) note, corrected P-values result in a loss of statistical power, especially when there are higher levels of population structure.

More recently, structured association has been the method of choice to correct for population structure in most association studies. For structured association, random unlinked markers are used to calculate and assign individuals into population substructures (Pritchard 2000). The program STRUCTURE (Pritchard 2000) is often used to calculate population structure. This program uses a MCMC Bayesian algorithm to calculate the proportion of an individual's genome that originated from different inferred populations. The individuals are then clustered into different groups based on their genome characterization. STRUCTURE

assumes that all individuals are unrelated and come from populations in Hardy-Weinberg equilibrium. STRUCTURE allows users to calculate the degree of population admixture of each individual. Principle component analysis (PCA) has also been used to calculate population structure (Price et al. 2006). This method can be much quicker than using STRUCTURE, and has been suggested by Zhao et al. (2007) to be just as effective.

There are two different types of models that apply structured association. The first is a GLM model which uses the subpopulations (Q) as covariates in a regression model, and then correlates the genotype with phenotype (Thornsberry et al. 2001). However, this model along with the GC model may not control false positives or have a low statistical power due to familial relatedness (Yu et al. 2005). The Q+K, or unified mixed model, still assigns subpopulations (Q) as covariates, but it also uses a kinship matrix (K) as a covariate in the regression (Yu et al. 2005). This method accounts for both population structure and familial relatedness. A number of studies have demonstrated that the Q+K model can be more effective than just the Q model. For example in Arabidopsis, Zhao et al. (2007) found that when cumulative P-values were plotted for flowering time, the Q+K model corrected for more false associations than the Q model. As a result, the Q+K model is a popular choice in most GWAS.

Erskine et al. (1989) highlighted that there are morphological differences between lentil accessions based on their regional adaptation. Accessions from Syria, Jordan, Egypt and Lebanon formed the Levantine group. Accessions from Europe, Turkey and Iran formed the northern group. Individuals from South Asia formed the Indian group, while individuals from Ethiopia formed the Ethiopian group. The groups were formed based on differences in days to maturity, pod set, and seed weight. These differences should result in changes in allele composition among the various groups. Therefore, it is expected that lentil will exhibit high levels of population sub-structure. In a previous population structure study of lentil, Liu et al. (2008) found there were eight subpopulations, using the program STRUCTURE. The individuals that were used came from 440 accessions located in the Chinese National Gene Bank. The eight clusters found in that study do not imply that all lentil

populations will cluster into one of these eight groups. For every association panel that is assembled, the population structure needs to be calculated. This is because there may be differences in adaptation or in pedigrees that could cause differing levels of population structure. In association panels that consist of cultivars or breeding material that share similar pedigrees kinship amongst the lines may also be high. This would make determining the population structure (Q) and kinship (K) necessary for GWAS in lentil.

2.11.5 Linkage and Association Mapping

Most studies implement linkage mapping and AM separately. Both methods can also be used to cross validate results from either of the studies. For example, linkage mapping was used to determine that loci which were significant for lung tumor susceptibility in inbred mice, were actually spurious (Manenti et al. 2009). But they were also able to confirm that the locus *Pas1* was significant for tumor susceptibility. The authors even proposed that linkage mapping should always be used in conjunction with association mapping studies in inbred mice. Brachi et al. (2010) also highlighted that dual mapping strategies can identify the false negatives that would not have been found if only one strategy was used when they analyzed flowering time in Arabidopsis. False negatives are loci that are actually associated with the phenotypic trait, but due to low frequency or population structure have low significance. In fact, their research showed that there was a 40% difference in the number of candidate genes identified to be false negatives when linkage and association mapping strategies were compared. Association mapping relies on the frequency of the alleles that exist within a panel to determine if there is a significant effect on a trait. If an allele is rare however, but still has a large effect, then its statistical power is weak and that allele would not appear to be significant. If that individual were to be used as a parent in a mapping population, and it segregated with other more common alleles, then this genomic region could be detected. Thus, when used together these methods can account for the limitations of the other and present a better model for the genetic control of a quantitative trait.

Chapter 3

Construction of a Genetic Linkage Map

3.1 Introduction

Genetic maps are derived from genotyping segregating populations, such as F_2's and RILs, with molecular markers and determining the level of recombination amongst the markers within each line. Genetic maps are needed for: QTL analysis, fine mapping, gene-based cloning, determining linkage disequilibrium, constructing physical maps, comparative genomics and a general understanding of the genetic architecture of a crop (Diaz et al. 2011).

Single nucleotide polymorphic (SNP) markers are abundant, distributed throughout the genome and are inherited co-dominatingly. Recently, a 1536 SNP Illumina Golden Gate SNP Assay (Lc1536) has been developed for lentil (Sharpe et al. 2013). The objective of this project was to genotype the intraspecific RIL population, LR-18, with the SNP assay and develop a linkage map that could be used for further QTL analysis.

3.2 Materials and Methods

3.2.1 Plant Material

LR-18 is a RIL population developed from a cross between the cultivar CDC Robin (Vandenberg et al. 2002) and the breeding line 964a-46. The F_2 population was advanced to the F_7 generation using single seed descent. Single F_8 plants were bulked to form the RILs. In total, 139 RILs were phenotyped and genotyped.

3.2.2 DNA Extraction and SNP Genotyping

Leaf samples were collected from multiple plants of each of the 139 RILs. Samples were freeze dried and stored at -80°C. The DNA was extracted using the micro-CTAB protocol (Doyle and Doyle 1990).

The Lc-1536 Golden Gate SNP OPA was used to genotype the population. Approximately 50ng of genomic DNA per sample was used for the assay. The assay involved amplification of DNA using fluorescently labelled allele-specific binding primers in a PCR reaction. The PCR products were then hybridized to beads. The

levels of fluorescence were then analyzed using the BeadStudio software (Illumina, San Diego, CA). Each sample was characterized by the ratio of the two allele scores that were determined by the level of fluorescence measured by the bead reading software. The program GenomeStudio ver 2010.1 (Illumina, San Diego, CA) was used to analyze and assign genotypes at each SNP locus to each sample (Figure 3.1). All genotyping information, including the map, is accessible through the University of Saskatchewan's Pulse Crop and Breeding group's web portal KnowPulse (http://knowpulse2.usask.ca/portal/).

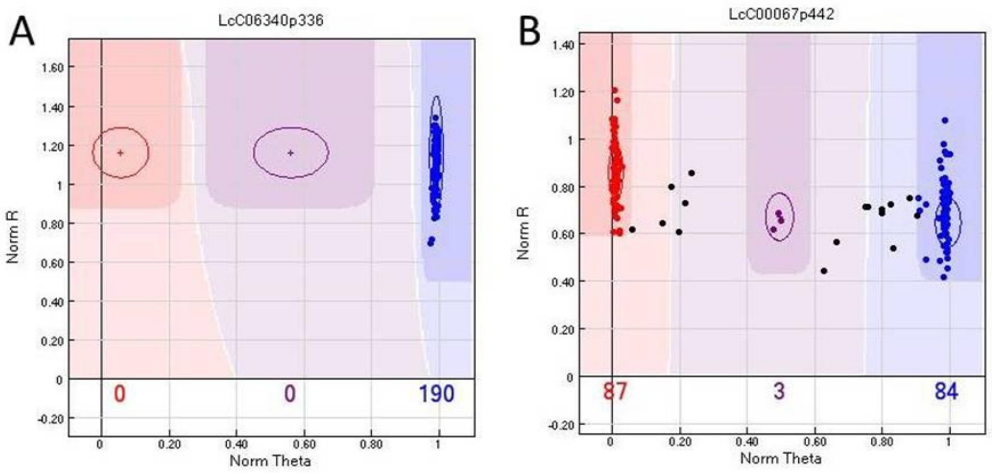

Figure 3.1. Example of SNP analysis using GenomeStudio ver 2010.1. The coloured dots correspond to each individual in the population. Monomorphic markers (A) appear clustered in the same region (blue), while polymorphic markers (B) show 1:1 segregation for each allele (blue and red). Heterozygous markers are located in between the two groups (pink). Markers not belonging to a distinct group are not coloured (black).

Other markers were genotyped on the population. There were 36 polymorphic SNPs genotyped using KASP assays. A total of 13 SSRs were also included. The population was also phenotyped for 4 morphological traits: cotyledon color (red or yellow: *Yc*), seed coat pattern (present or not: *scp*), and seed coat ground colour (brown, grey, tan and green: *Ggc* and *Tgc*).

3.2.3 Linkage Map Construction

A linkage map was developed using the program JoinMap 4.0 (Van Ooijen and Voorips 2004). A minimum logarithm of odds (LOD) value of 6 was used to

assign groups using the tree command. The maximum likelihood method (MLM) was originally used when calculating the linkage groups. Regression mapping, using the Kosambi mapping function, was used to finalize the map order and distances between the markers.

To confirm the linkage groups, each contig sequence from each marker was aligned with the *Medicago* genome to form a dotplot (Figure 3.2.). This resulted in the lentil linkage groups matching with *Medicago* linkage groups. There are a number of inversions and translocations that distinguished lentil from *Medicago*, however the collinearity between the linkage groups was sufficient to confirm they were homologous with the *Medicago* chromosomes. Linkage groups 1 through 5 aligned to *Medicago* chromosomes 1 through 5. Linkage group 6 aligned to chromosome 7 and linkage group 7 aligned to chromosome 8.

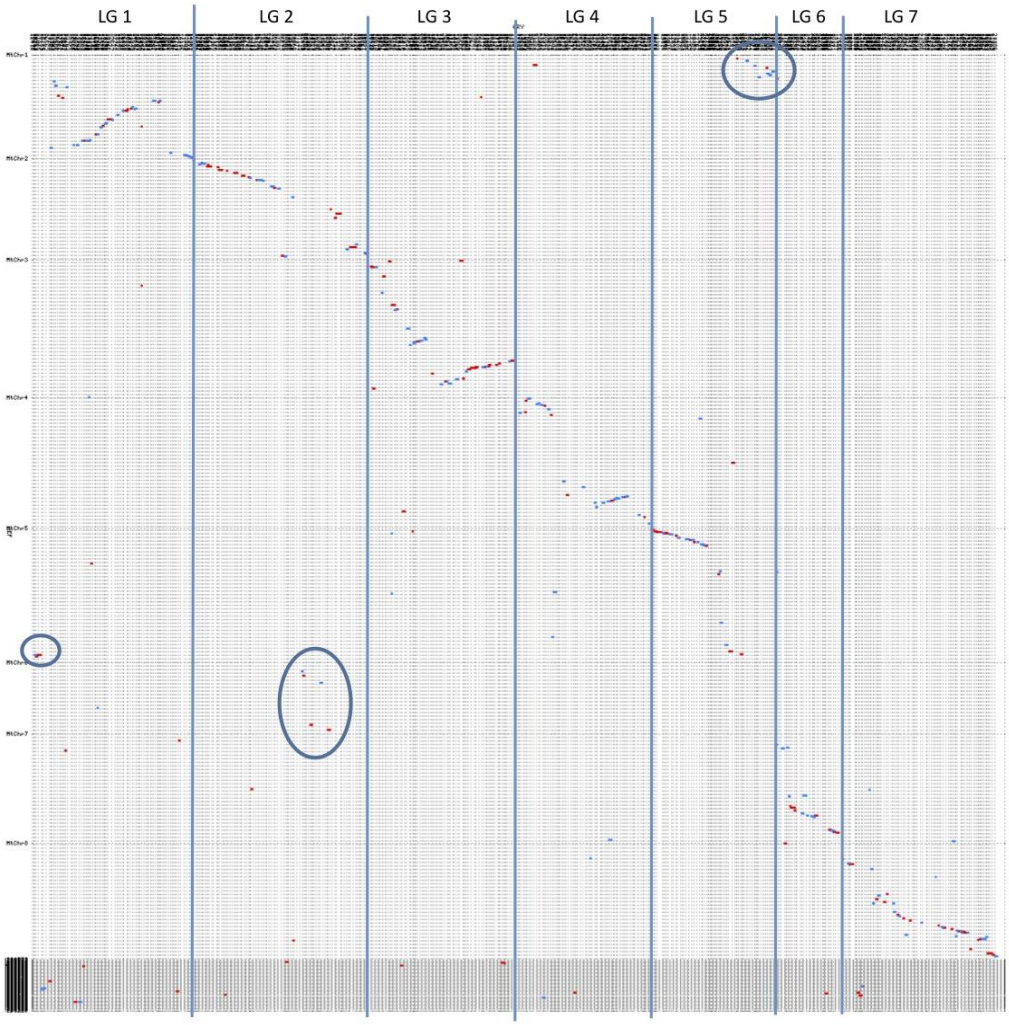

Figure 3.2. Dotplot showing collinearity between lentil and *Medicago*. The marked circles represent translocations. Linkage groups for lentil were selected based on their match with the chromosomes in *Medicago*. (reproduced with permission from Sharpe et al. 2013)

3.3 Results and Discussion

A 1536 Illumina Golden Gate assay and a KASP assay were used to screen the population for SNP markers. A total of 579 polymorphic SNPs were used in the development of the map (Table 3.1). Of the included SNPs, 17 were dominant markers. Thirteen SSRs and 4 morphological markers were also included. A total of 561 markers were mapped to seven linkage groups (Figure 3.3.) Eighteen markers remained unmapped.

Table 3.1. Markers used for the construction of the lentil linkage map.

Marker Type	Number of polymorphic markers	Number of mapped markers
SNPs	562	547
SSRs	13	10
Morphological	4	4
Total	579	561

The map developed in this study is the most condensed linkage map in lentil to date. A total of 561 markers were mapped with a total length of 597 cM. This map has an average distance of 1.06 cM between markers. The physical size of the lentil genome was estimated to be 4,086 Mbp (Arumuganathan and Earle 1991). Based on the size of the map developed in this study, each cM represented an average of 6.8 Mbp. The Tullu et al. (2008) intraspecific map covered 1868 cM with an average distance of 8.7 cM between markers. The physical to genetic distance ratio of that map was 2.18 Mbp/cM. Kahraman et al. (2004) developed a map spanning 1192 cM with an average marker distance of 9.1cM. Rubeena et al.'s (2003) linkage map had a total distance of 784.1 cM with an average distance of 6.8cM. Phan et al. (2009) produced a map with a total distance of 928.4 cM and with an average distance of 9.5 cM. Compared to consensus maps in other crops, such as barley, which have an average distance of 0.38 cM between markers (Muñoz-Amatriaín et al. 2011), lentil could still benefit from an increased number of mapped markers.

The length of each linkage group varied. Linkage group 2 was the longest with a total length of 150 cM, whereas linkage group 7 was the shortest with a length of 57 cM. There have been other linkage studies in lentil that have been able to form seven linkage groups, but these maps were generated from interspecific populations (Eujayal et al. 1998). Prior to this study, intraspecific maps of lentil have not been able to be resolved into seven linkage groups. Rubeena et al. (2003) were able to generate 9 linkage groups. Tanyolac et al. (2010) and Saha et al. (2010) constructed maps with 11 and 12 linkage groups, respectively. Even more recent studies, such as Gupta et al. (2012), have only been able to resolve 11 linkage

groups. The differences seen amongst these linkage maps in lentil could be due a number of differences. The types of markers, the parents of the population and even the LOD scores can impact the size and number of linkage groups (Tullu et al. 2008, Paran et al. 1995).

Comparative genome mapping with model species is proving to be a useful method to determine linkage group numbers. By measuring the collinearity with the model species *Medicago*, linkage groups can be selected or combined to match with homologous *Medicago* chromosomes. Phan et al. (2007) was the first to use this method in lentil, and was successful in resolving 7 linkage groups. Using an even greater number of gene-based markers, this study found high levels of linkage group collinearity between lentil and *Medicago,* which helped in the formation of the 7 linkage groups.

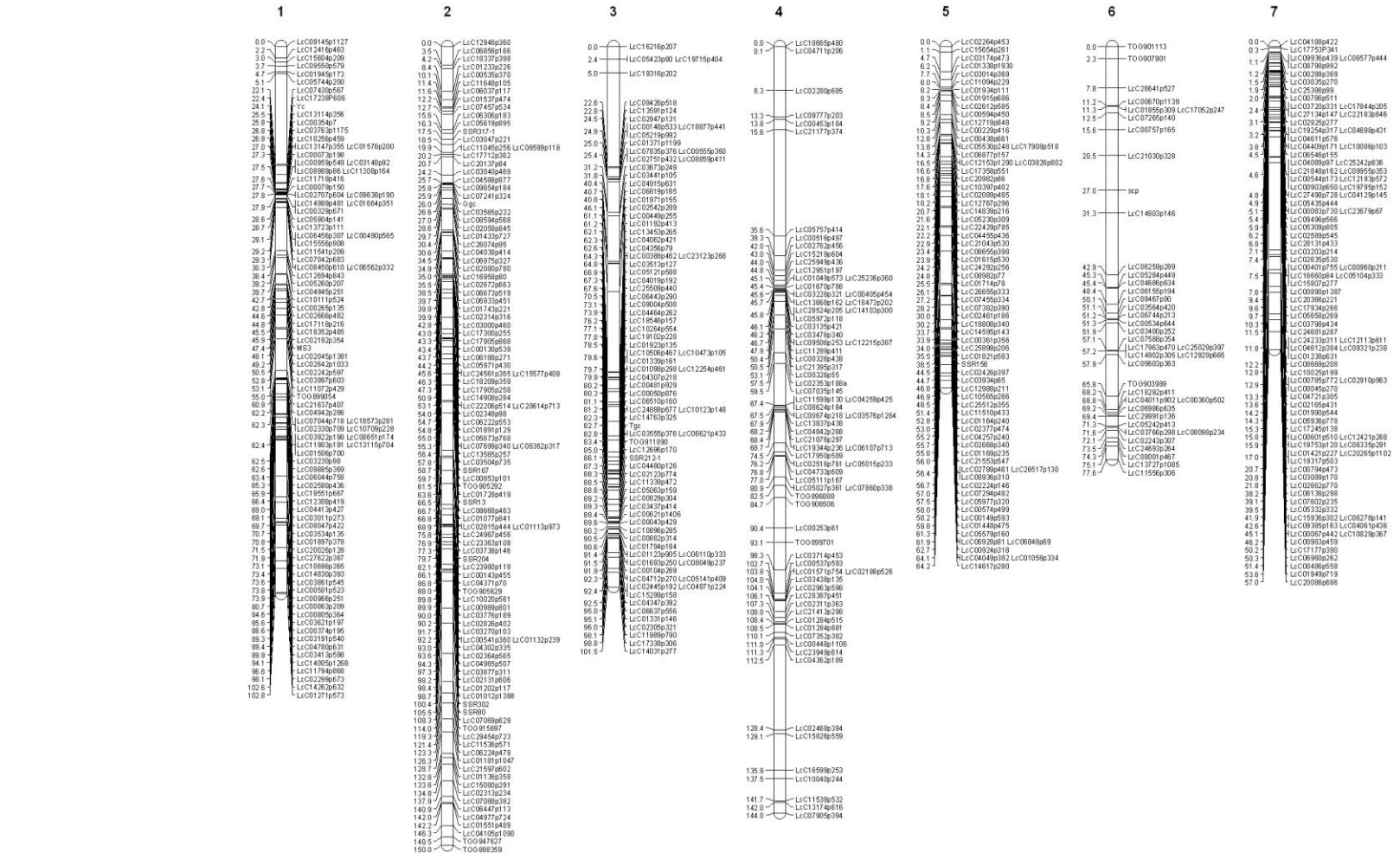

Figure 3.3 Linkage map of the LR-18 population. Seven linkage groups are shown which correspond to the seven chromosomes of lentil.

Segregation distortion is caused by regions in a chromosome that do not transmit equally to the progeny (Thoquet et al. 2002). Chi-square analysis was used to detect any marker segregation distortion in this population. A segregation ratio of 1:1 at P>0.05 was expected. In total 157 markers, or 28% of all mapped markers, showed significant distortion. Linkage group 2 showed the highest number of distorted loci with 59% of the markers not meeting the expected ratio. Linkage group 6 and 7 had the lowest levels of linkage distortion with 10% of the markers showing distortion. Other linkage maps in lentil have reported similar levels of linkage distortion. Saha et al. (2010) noted that 22% of the markers used in their linkage map, comprised of SSR, RAPD, SRAP and morphological markers, showed distortion. However, other maps have reported lower levels. The intraspecific lentil map published by Rubeena et al. (2003) contained RAPDs, ISSRs, and resistant gene analogs (RGAs) and had a 14.4% marker distortion. Eujayl et al. (1998) noted segregation distortion to be only 8.4% in their interspecific lentil map. Interspecific populations normally have greater levels of segregation distortion because of the lack of homology between the chromosomes of the different species (Flandez-Galvez et al. 2003). It is interesting to note that an interspecific population would have a lower segregation distortion versus the intraspecific population used in this study. However, for the 1536 Golden Gate assay, it has been noted that some of the SNP markers segregated in a way that represented gene duplications (Sharpe et al. 2013). Markers that are located within sequence duplications may not segregate in the expected ratio of a single gene, which would then appear as segregation distortion (Xian-Liang et al. 2006). This is a possible explanation for the higher levels of segregation distortion.

Within this LR-18 map there were regions that showed clusters of SNPs closely mapped together. The clusters were then separated by large distances (>10cM). This might be explained by how the SNP markers were developed. They were selected based on 3'-cDNA transcript profiling, using 454 sequencing, of multiple lentil genotypes. Therefore, all the SNPs came from only the coding regions of lentil and none came from the non-coding regions. King (2002) noted that

organisms with large genomes contain clusters of genes separated by non-coding regions like repetitive DNA. This may account for the clustering observed in lentil.

For this study we developed a linkage map from the cultivated gene pool in lentil. The parents of the population are two elite lines; CDC Robin is a cultivar and 964a-46 is a breeding line. Maps developed from better adapted germplasm will provide a more applicable map because the markers mapped are polymorphic within the gene pool. Breeders should find this map useful as many of the markers will also be polymorphic within their elite breeding material, thus making this map suitable for mapping QTLs.

In conclusion, we have developed the first linkage map in lentil using SNP markers. A version of this map, without the morphological markers, has been published (Sharpe et al. 2013). The use of SNP markers has greatly increased the number of mapped markers in lentil. It will also increase the resolution of QTL mapping and provide more robust markers for MAS. This map could also be used for improved comparative genomic analysis with other legume species, such as *Medicago truncatula*, which could lead to candidate gene identification.

Chapter 4

Linkage Mapping of Seed Size and Shape QTLs

4.1 Introduction and Objectives

Quantitative traits have been mapped in lentil for the purpose of associating molecular markers with phenotypic traits. This would help accelerate crop breeding by means of MAS. However, in lentil very few molecular markers are used in MAS for breeding. This is primarily because many of the molecular markers are not reproducible in the breeding germplasm, or are difficult and too expensive to screen (Vail 2010; Ford et al. 2009). Recently, Sharpe et al. (2013) developed a lentil 1536 Golden Gate SNP assay that is reproducible and locus specific in the *Lens culinaris* ssp. *culinaris* genetic background. In Chapter 3, a linkage map was developed using this assay in the LR-18 population. We can now use the linkage map to map quantitative traits, like seed size and shape.

The purpose of this study was to detect the genomic regions associated with seed size, seed shape and flowering time in a mapping population, using SNP markers, evaluated at two different locations in Saskatchewan, Canada. The objective was to produce functional SNP markers associated with those traits which could be used for routine MAS in the breeding program.

4.2 Materials and Methods

4.2.1 Plant Material

Lentil recombinant inbred line (RIL) population, LR-18, was developed from a cross between cv. CDC Robin (Vandenberg et al. 2002) and the line 964a-46 from the lentil breeding program at the Crop Development Centre, University of Saskatchewan. CDC Robin produces seeds that are small in diameter but relatively plump. The breeding line 964a-46 produces seeds that are large in diameter but not as plump (Figure 4.1).

A total of 147 F_7-derived RILs were assessed at two different locations: Saskatoon (Preston) and 15 km SE of Saskatoon (SPG) in 2009 and 2011. The RILs

were grown in $1m^2$ microplots in a randomized complete block design with 3-replicates.

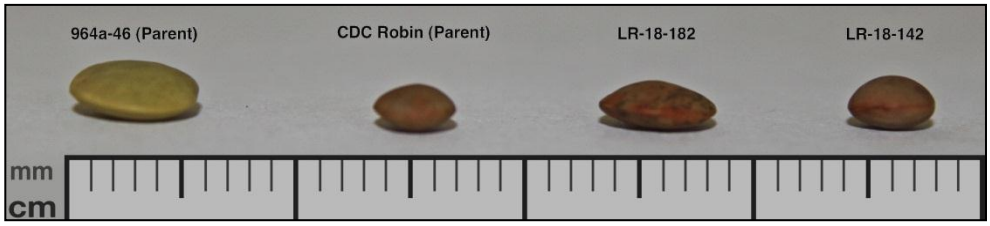

Figure 4.1. Differences in seed size and shape among the parents and two randomly chosen RILs from the LR-18 population.

4.2.2 Phenotyping

The number of days to 50% flowering (DTF) was recorded for all plots at all locations. The harvested seed samples were measured for seed diameter and seed thickness using round-hole and slotted sieves, respectively, as described by Hossain et al. (2010). Seed diameter was measured by passing at least 50g of a sample through a set of seven round-holed sieves from 5.8 mm (15/64") down to 3.6 mm (9/64") in 0.25 mm (1/64") increments. Thickness was measured by passing the same samples through a set of six slotted sieves from 2.8 mm (7.5/64") down to 2.0 mm (5/64") in 0.2 mm (0.5/64") increments. All samples were shaken through the sieves for one minute on a flat-bed shaker prior to weighing the seed retained on each sieve. Values for seed diameter and thickness for each sample were then calculated using the formula:

$$\% \text{ on sieve} = \text{wt(g) on sieve}/\text{wt(g) total sample} * 100$$
$$\text{Seed Size} = \Sigma(\% \text{ on sieve} * \text{sieve hole size (mm)})/100$$

Seed plumpness was calculated by dividing the seed thickness by the seed diameter values for each genotype.

To determine the accuracy of this method a comparison with caliper-based data was made. For each of 25 samples, 10 seeds were randomly selected and measured for seed diameter and seed thickness with a caliper. The means were then calculated from the ten measurements. Plumpness was determined by dividing

thickness by diameter. The values were then correlated with the values developed by the seed screening method.

Cotyledon colour (*Yc* - yellow or red; Slinkard 1978) was determined for each of the RILs. Seed coat colour was used to determine which allele at *Ggc* and *Tgc* they were carrying as well as if there was a seed coat pattern or not (*Scp*) (Vandenberg and Slinkard 1990).

4.2.3 Statistical Analysis

The years and locations of the field trials were combined to form site-years. All statistical analyses were done using the software R v2.11.1 (R Development Core Team 2011). A linear mixed-model was fit using the reps and site-years as random factors, while the genotypes were considered fixed. The R package nlme was used to fit the linear mixed model using the lme function (Pinheiro and Bates 2000). The R package lme4 was used to calculate the variance components under a mixed model using the function lmer (Bates 2007).

4.2.4 QTL Analysis

A linkage map, consisting of 547 SNPs, 10 SSRs (Sharpe et al. 2013) and the four morphological markers (*Yc, Ggc, Tgc* and *Scp*) was constructed for the LR-18 population (Chapter 3) and used for QTL analysis. QTL analysis was done using MapQTL 5.0 (Van Ooijen 2004). One thousand permutation tests were run to determine the LOD threshold value. A value of 3.0 was determined and used to declare significant QTLs. Interval mapping (IM) was used for each location and year. Markers that showed high LOD values were selected as co-factors and through composite interval mapping (CIM) were analyzed for QTLs.

4.3 Results

4.3.1 Phenotypic Data

The seed diameter, thickness and plumpness values that were determined using the sieve screening method were significantly ($p<0.05$) correlated with the values measured with a caliper. The estimates of seed diameter ($r=0.90$), seed thickness ($r=0.88$) and seed plumpness ($r=0.91$) all had high correlations. This gave

confidence in the use of the sieving method for assigning seed morphology values to each sample.

Analysis of variance revealed that genotype had a significant effect (P≤ 0.001) on all seed morphology traits and flowering time (Table 4.1). Site-year was significant for seed thickness, seed plumpness and flowering time. The genotype x site-year interaction was also significant for all traits.

Table 4.1. ANOVA results including F-values for seed diameter, thickness, plumpness and DTF for genotype and environmental effects.

		F – values			
Effect	df	Diameter	Thickness	Plumpness	Flowering Time
Genotype	146	114.0***	15.5***	150.8***	8.26***
Site-Years	3	2.9 ns	165.2***	445.5***	468.04***
Genotype x Site-Years	438	2.1***	1.9***	2.6***	1.96***
CV%		8.11	5.05	8.55	8.54

*** significant at P≤ 0.001, ns not significant

The overall means of seed thickness and seed plumpness in the 2009 growing season were greater than in 2011, while for seed diameter there was no difference between the years (Figure 4.2). The 2009 SPG site had the greatest seed thickness and also the greatest seed plumpness versus the other site-years. DTF showed the most site-year variability amongst all the traits measured. The SPG site had longer DTF for both 2009 and 2011 compared to the Preston site (Figure 4.2).

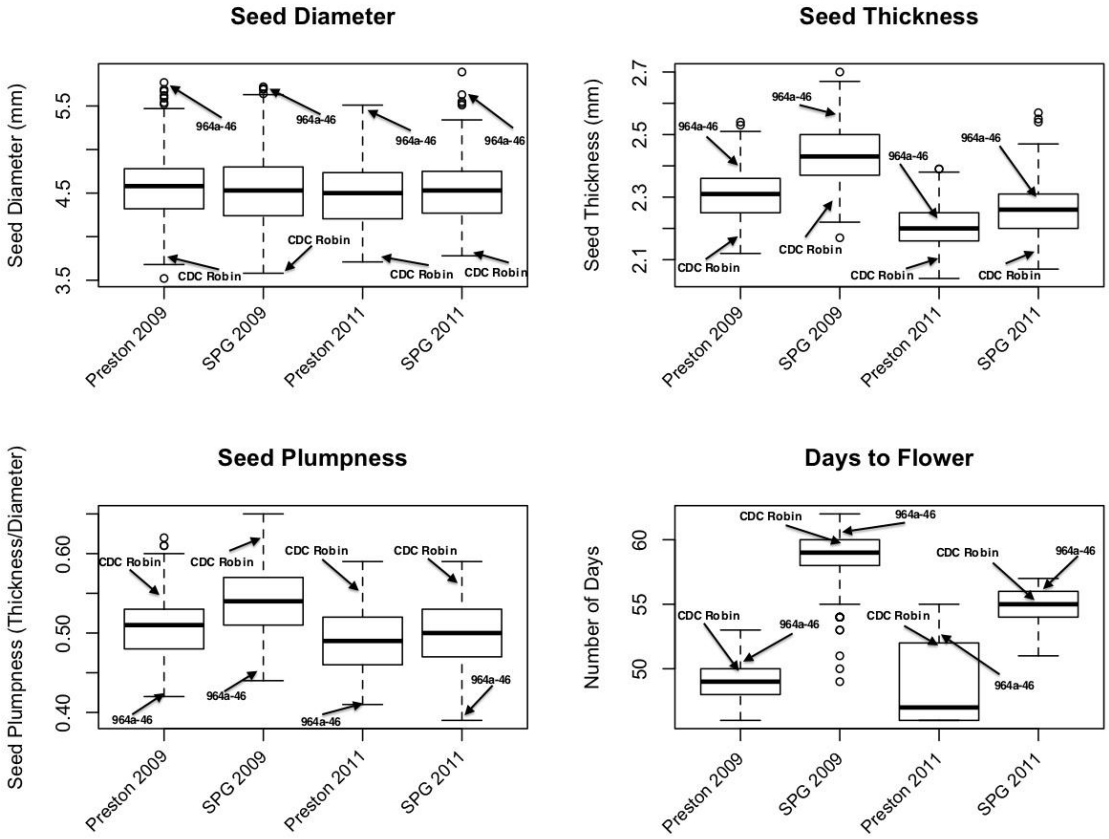

Figure 4.2. Box and whisker plots of the distribution of seed diameter, thickness and plumpness, and DTF in the LR-18 (CDC Robin x 964a-46) RIL population grown at Preston and SPG in 2009 and 2011. The mean value of the parents are labeled with an arrow.

Pearson's correlation (r) coefficients were calculated for all the traits measured within each of the site-years (Table 4.2). All traits showed some level of significant correlation, but for some traits certain site-years were not significant. Seed diameter and seed plumpness showed the highest significant correlation with all site-years averaging a negative correlation of r=-0.90. All site-years for seed diameter and seed thickness correlations were significant but remained below r=0.40. Only two of the four site-years had a significant correlation between seed thickness and seed plumpness. DTF was significantly correlated with all seed

morphology traits for all site-years. DTF had the highest correlation with seed plumpness, averaging r=0.36.

Table 4.2. Pearson's correlation coefficients of seed traits and DTF, for each site-year

Trait	Site Year	Diameter	Thickness	Plumpness	DTF
Thickness	SPG 2009	0.38***	-		
	Preston 2009	0.16***	-		
	SPG 2011	0.22***	-		
	Preston 2011	0.32***	-		
Plumpness	SPG 2009	-0.90***	0.04ns	-	
	Preston 2009	-0.90***	0.28***	-	
	SPG 2011	-0.88***	0.26***	-	
	Preston 2011	-0.92***	0.04ns	-	
DTF	SPG 2009	0.17***	-0.12*	-0.24***	-
	Preston 2009	0.21***	-0.30***	-0.34***	-
	SPG 2011	0.40***	-0.21***	-0.50***	-
	Preston 2011	0.26***	-0.17**	-0.34***	-

*** significant at P≤ 0.001, ** significant at P≤ 0.01, * significant at P≤ 0.05, ns not significant

Variance components were used to calculate the heritability of each trait (Table 4.3). Seed diameter and plumpness were highly heritable (0.92 and 0.94, respectively); while seed thickness had a more moderate heritability (0.60) and the heritability of DTF was even lower (0.45).

Table 4.3. Estimates of variance components and broad-sense heritability for seed size traits and DTF for RILs grown at two locations over two years (four site-years).

Variance Component	Diameter	Thickness	Plumpness	DTF
σ^2G	0.12	0.003	0.0015	1.14
σ^2G * Site Year	0.0024	0.0007	0.00006	0.61
σ^2E	0.0065	0.003	0.00001	2.13
σ^2P	0.13	0.005	0.0016	2.52
H^2	0.92	0.60	0.94	0.45

4.3.2 QTL Analysis

QTLs were located on five of the seven linkage groups (Table 4.4, Figure 4.3). For seed diameter, three different QTLs were identified, all of them present in all site years. The QTL that accounted for the most variation (>23%) was located near the cotyledon colour locus (*Yc*) on linkage group 1. The other major QTLs for seed diameter were located near the SNP markers LcC00853p101and LcC00890p1387 on linkage groups 2 and 7, respectively. Together these three QTLs explained at least 60% of the variation for seed diameter for all site-years. The additive effects results indicated that the seed diameter allele for each of these markers came from the large seed diameter parent, 964a-46 (Table 4.4).

Seed thickness QTLs were detected on linkage groups 1,2,4,5,6 and 7. There were multiple QTLs that were specific to each site year. The QTL that was most stable throughout the different site-years was the one located on linkage group 7 (Figure 4.3). This QTL explained an average 8.4% of the variation in each of the three site years it was declared significant. The additive effects also showed that the allele contributing to the QTL came from CDC Robin.

Seed plumpness QTLs were present on linkage groups 1, 2 and 7. Seed plumpness shared the same QTL on linkage group 1 with seed diameter at the cotyledon colour (*Yc*) locus. The QTL present on linkage group 7 also shared the same marker locus (LcC00890p1387) with the seed diameter QTL. The third QTL, linked to the SNP marker LcC00853p101, mapped to the same location as the seed diameter QTL on linkage group 2. Both the QTLs on linkage groups 1 and 7 explained the majority of the variation with their combined values explaining over 50% of the variation for each site-year. The QTL located on linkage group 2 explained less than 10% of the variance.

QTLs identified for DTF were located on linkage groups 1, 2 and 7. The QTL located on linkage group 1 was the only QTL for DTF that was significant in multiple site-years. This QTL was also located in the same genomic region as *Yc* and the seed diameter and seed plumpness QTLs.

Table 4.4 QTLs identified for seed diameter, seed thickness, seed plumpness and DTF at four site years.

Trait ψ	Site †	Marker	Linkage Group	Position	LOD	% Exp	Add. Effects
Diameter	Pres 09	*Yc*	1	23cM	26	31	-0.18742
Diameter	SPG 09	*Yc*	1	23cM	23	26.9	-0.2059
Diameter	Pres 11	*Yc*	1	23cM	30.5	38.4	-0.22967
Diameter	SPG 11	*Yc*	1	23cM	29	36.3	-0.20761
Diameter	Pres 09	LcC00890p1387	7	7.5cM	21.3	22.7	-0.15645
Diameter	SPG 09	LcC00890p1387	7	7.5cM	24.3	29.6	-0.21072
Diameter	Pres 11	LcC00890p1387	7	7.5cM	21.4	23.1	-0.17403
Diameter	SPG 11	LcC00890p1387	7	7.5cM	21.6	24.2	-0.16489
Diameter	Pres 09	LcC00853p101	2	59.7cM	15.6	14.8	-0.13508
Diameter	SPG 09	LcC00853p101	2	59.7cM	11.8	10.7	-0.13533
Diameter	Pres 11	LcC00853p101	2	59.7cM	9.9	8.5	-0.11308
Diameter	SPG 11	LcC00853p101	2	59.7cM	10.5	9	-0.10783
Thickness	Pres 09	LcC04409p171	7	3.7cM	12.35	25.5	0.033645
Thickness	Pres 11	LcC04409p171	7	3.7cM	4.45	11.7	0.016051
Thickness	SPG 11	LcC04409p171	7	3.7cM	8.51	23.5	0.035911
Thickness	Pres 09	LcC09777p203	4	13.3cM	4.01	7.8	-0.01811
Thickness	Pres 11	LcC09777p203	4	13.3cM	3.08	7.2	-0.0127
Thickness	Pres 09	LcC05284p449	6	45.3cM	3.55	6.4	-0.01686
Thickness	SPG 09	LcC05284p449	6	45.3cM	4.01	11.1	-0.02767
Thickness	SPG 09	LcC05579p160	5	61.25cM	3.04	8.3	-0.02399
Thickness	Pres 11	LcC02348p98	2	53.9cM	4.56	12.1	-0.01706
Thickness	Pres 11	LcC20026p128	1	71.4cM	3.86	10.1	-0.01517
Thickness	SPG 11	LcC05332p332	7	39.4cM	3.59	9.2	-0.02246
Plumpness	Pres 09	*Yc*	1	23cM	23.85	27.6	0.021228
Plumpness	SPG 09	*Yc*	1	23cM	20.86	22.2	0.021197
Plumpness	Pres 11	*Yc*	1	23cM	34.2	39.8	0.024883
Plumpness	SPG 11	*Yc*	1	23cM	28.28	32.4	0.022619
Plumpness	Pres 09	LcC00890p1387	7	7.5cM	27.98	35.5	0.023445
Plumpness	SPG 09	LcC00890p1387	7	7.5cM	26.79	32.4	0.025598
Plumpness	Pres 11	LcC00890p1387	7	7.5cM	28.64	30.4	0.021251
Plumpness	SPG 11	LcC00890p1387	7	7.5cM	29.71	35.6	0.022976
Plumpness	Pres 09	LcC00853p101	2	59.7cM	6.69	5.6	0.010055
Plumpness	SPG 09	LcC00853p101	2	59.7cM	6.47	5.2	0.010969
Plumpness	Pres 11	LcC00853p101	2	59.7cM	5.5	3.6	0.007987
Plumpness	SPG 11	LcC00853p101	2	59.7cM	5.3	4	0.008383
DTF	Pres 09	*Yc*	1	23cM	5.23	12.6	-0.43309
DTF	Pres 09	LcC06044p758	1	63.4cM	4.55	10.9	-0.39292
DTF	SPG 09	*Yc*	1	23cM	3.58	10.8	-0.50231
DTF	Pres 11	*Yc*	1	23cM	8.16	20.7	-1.17264

DTF	Pres 11	LcC09496p566	7	5.4cM	5.23	13	-0.90865
DTF	SPG 11	*Yc*	1	23cM	15.86	34.9	-0.58147
DTF	SPG 11	LcC23363p108	2	76.9cM	3.48	6.2	-0.23647

Ψ Diameter: Seed diameter; Thickness: Seed thickness; Plumpness: Seed Plumpness; DTF: Days to Flowering

† Pres: Preston; SPG: Sask Pulse Growers

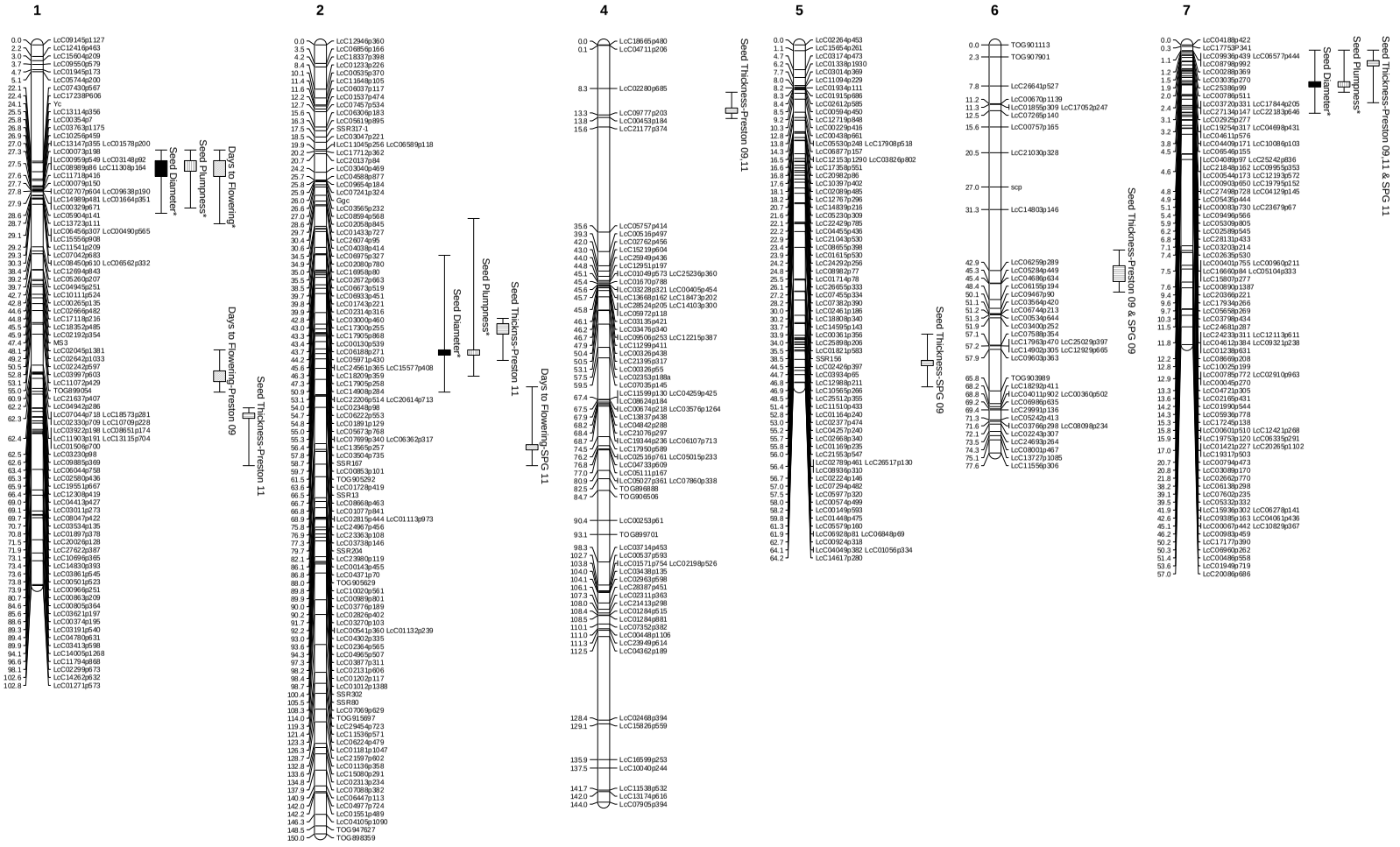

Figure 4.3. Genetic linkage map of lentil RIL population LR-18 with QTLs for seed diameter, seed thickness, seed plumpness and DTF. All linkage groups are shown with QTLs marked next to the loci they are associated with. The thin lines represent the regions that were significant using interval mapping, while the boxes represent the regions significant under composite interval mapping. QTL boxes marked with an asterisk (*) represent QTLs that were significant for all site-years

4.4 Discussion

In this study, lentil seed diameter, seed thickness and seed plumpness, along with DTF, were phenotyped in the mapping population LR-18 and then analyzed for QTLs using the linkage map developed in Chapter 3. This is the first study to map seed size and shape QTLs in an intraspecific lentil population. This is also the first reported study in lentil using SNP markers for QTL mapping.

For this study seed weight was not measured. Seed weight could be used to differentiate large versus small seeds, but would not be practical in determining seed shape. For example in rice, QTLs for seed weight and QTLs for seed length, width and length/width were not associated with one another (Qiu et al. 2012). Furthermore, Fratini et al. (2007) investigated seed weight and seed diameter in an intraspecific sub-species lentil population, and found a significant but low (0.34) correlation between the two traits. Fratini et al. (2007) also mapped both seed weight and seed diameter QTLs and found three different QTLs for both traits. None of those QTLs were co-located. This suggests that seed weight, or seed weight marker loci, would not be suitable for selecting for size and shape traits, such as seed diameter and seed plumpness, in lentil.

Studies in other crops have revealed that seed size is a complex trait that is highly influenced by the environment (Cobos et al. 2007; Breseghello and Sorrells 2007). In lentil, previous studies involving seed size QTLs (Fratini et al. 2007; Abbo et al. 1991) have not addressed environmental interactions. In this current study, significant genotype x environmental interactions were detected. Throughout lentil seed development, seed diameter appears to reach its maximum size quickly and is more buffered against variability in the environment. Seed thickness however, appears to be more subject to environmental variability, with ideal growing conditions resulting in thicker seed. The 2009 and 2011 growing seasons differed considerably in Saskatchewan. The mean temperature for the months of May to August in 2009 was 13.8°C and 15.5°C in 2011 (Environment Canada 2011). The total precipitation from May to August in 2009 and 2011 were 215(mm) and 197(mm), respectively. Interestingly, the 2009 growing year had nearly 50(mm) greater precipitation compared to 2011 for the months of July and August, the

months of flowering and seed filling. The greater moisture availability after flowering time, in the 2009 growing season may explain why seeds were thicker and plumper in that year. Seed diameter, however, showed no significant differences in variability across the different environmental conditions. No studies in lentil have mentioned the underlying mechanisms that determine why certain traits in lentil seed development are more susceptible to environmental differences. However, there are a number of studies in other legume crop species where seed development has been closely examined (Le et al. 2007). Domoney et al. (2006) also highlighted that there are largely two distinct phases in legume seed development of which seed diameter, seed thickness and seed plumpness would belong to. The first developmental phase, cell division, is noted to depend on the embryo genotype, which controls the cotyledon cell number. This phase is largely insensitive to environmental variability. Due to low environmental interactions and high heritability, seed diameter and seed plumpness can be considered to be controlled by loci that are regulated in this developmental phase. The second phase, cell expansion, is highly influenced by the environment, and has been noted to be regulated by loci involved in photosynthate partitioning. Seed thickness would belong in this developmental phase. However, it is worth noting that there were genotypic effects for seed thickness, suggesting that it is not entirely controlled by environmental conditions. Loci that influence the rate of photosynthate accumulation in the seed could be contributing to this genetic variability.

The level of environmental variation for each trait was reflected in the heritability estimates. Both the diameter and plumpness heritabilities were very high, but the seed thickness was more moderate. Flowering time had the lowest heritability (0.45) which is probably a reflection of the noticeable difference between the average DTF from location to location and year to year. These results for flowering time agree with the findings of Tullu et al. (2008) who found that flowering time in their lentil population had an even lower heritability of 0.31. Results presented in this study showed much higher seed size and shape heritabilities relative to other legume crops. In soybean, Cober et al. (1997) found that heritability for seed size ranged from 0.26-0.50 and seed shape heritabilities

ranged from 0.59-0.75 among four populations. In common bean (*Phaseolus vulgaris*) the heritability for seed shape was 0.61 (Genchev 2006). However, both soybean and common bean have different morphological seed characteristics compared to lentil. Like many legume species, their seed shape is determined by a three axial dimension (length x width x thickness) instead of the two axial (diameter x thickness) ratio of lentil.

QTL analysis was used to dissect the quantitative nature of these traits and identify regions of the genome that contributed to the genetic variability. It was observed that all seeds with large diameter RILs had yellow cotyledons and the RILs with smaller diameter seeds had red cotyledons suggesting a high level of linkage between seed diameter and cotyledon color. It was not surprising, therefore, to discover that the seed diameter QTL that explains the highest level of variation was linked to the *Yc* locus. These findings agree with Abbo et al. (1991) who mapped seed weight QTLs to the cotyledon color locus in two interspecific populations, *L. culinaris* ssp. *culinaris* x *L. c.* ssp. *orientalis* and *L. culinaris* ssp. *culinaris x L. ervoides.* In chickpea, it has also been reported that seed weight and beta-carotene, a carotenoid which controls cotyledon color, also share the same QTL (Abbo et al. 2005). However, when Fratini et al. (2007) mapped the cotyledon colour marker in an F_2 population derived from a cross between *L. culinaris* ssp. *culinaris* and *L. c.* ssp. *orientalis,* they found no association between this locus and seed diameter QTLs. Tullu et al. (2001) also found that there was large variation in seed weight compared to cotyledon color when analyzing a lentil core collection. This suggests that the linkage between seed diameter and cotyledon colour may be specific to certain populations. Furthermore, within the lentil breeding program at the CDC there has been segregation of seed diameter and cotyledon colour noted in populations derived from crosses between yellow and red cotyledon parents (A. Vandenberg, pers. comm.). This material could be used for further breeding in developing improved seed diameter in lentil.

Seed thickness QTLs were detected on 6 of the 7 linkage groups. Only three of those QTLs were significant in multiple site years. This highlights the genotype by environment interactions for seed thickness. Seed thickness was significantly

correlated with seed diameter and DTF at all site-years. It was also correlated with seed plumpness, but for only two site-years. However, no QTLs for seed thickness were shared with the other seed morphology or flowering time QTLs. The marker LcC04409p17 did map close, ~4cM, to the marker LcC00890p1387 which was associated with seed diameter and seed plumpness QTLs.

For seed plumpness, all three QTLs reported were also located in the same position as seed diameter QTL. These three QTLs were located on linkage group 1 at the *Yc* locus, linkage group 2 at the SNP marker LcC00853p101 and linkage group 7 at the SNP marker LcC00890p1387. Sharing the same QTLs for seed diameter and seed plumpness was expected because the correlations between the two traits were high ranging from r=-0.88 to r=-0.92 and were significant for all site years. Salas et al. (2006) evaluated seed shape traits in soybean and found that there were certain QTL regions that controlled multiple seed traits like seed length, height, weight and volume. This suggests that certain seed quality/morphology traits in lentil, and in other legume crops, are inherited together, either through linkage or pleiotropy. This would make breeding for each trait, independent of other seed morphology traits, difficult. In retrospect, the LR-18 population may not have been ideal for studying seed plumpness in lentil. The high correlation between diameter and seed plumpness suggests that, in this population anyway, seed diameter highly influences the level of plumpness that a given genotype may have. It could be because the population is actually not segregating for seed plumpness. A possible solution would be to use a mapping population where the two parents have nearly the same seed diameter, and maybe belonging to the same market class, but differ in their seed plumpness.

DTF was significantly correlated with all the seed size and shape traits; albeit all correlation values were below 0.50 (Table 4.2). The only QTL significant in multiple site years for DTF was located on linkage group 1 at the locus *Yc,* the same as seed diameter and seed plumpness. In soybean and common bean, flowering time and seed size QTLs have also been mapped to co-incident regions (Watanabe et al. 2004; Pérez-Vega et al. 2010). Another example is the *AP2* gene in Arabidopsis, which is known to control flower development and seed size (Jofuku et al. 1994).

This suggests that the *Yc* locus could be controlled by a similar mechanism, where one gene controls multiple seed and flowering time traits. However, Sarker et al. (1999) mapped the *Yc* marker and flowering time in an interspecific lentil population and found them to be unlinked. Fratini et al. (2007) also mapped a flowering time QTL in an interspecific population and found no seed weight or seed diameter QTLs were nearby. The flowering time QTL in the Fratini et al. (2007) study mapped near the seed coat pattern (*scp*) locus, which mapped to linkage group 6 in the LR-18 population. No QTLs for flowering time were observed on linkage group 6 in the LR-18 population.

With the recent developments in lentil genomics, described by Sharpe et al. (2013), comparative analysis can be used to determine if any of these genes located in model species correspond to the QTLs identified in this study. With the availability of high throughput genome scans for lentil, association mapping can now be applied to diverse material. The higher levels of recombination contributing to the lower levels of linkage disequilibrium found in more diverse material could result in finer mapping of markers. This could also result in less co-inheritance of traits. Association mapping studies, in conjunction with this one, could lead to improved strategies for marker-assisted breeding to select for specific shapes and sizes of lentil.

Chapter 5

Association Mapping of Seed Size and Shape in Lentil

5.1 Introduction and Objectives

In Chapter 4 a bi-parental population was evaluated for seed diameter, seed thickness, seed plumpness and DTF QTLs. It was observed that many of the QTLs shared the same mapping positions and appeared to be inherited together. Observations throughout the lentil species shows that broad phenotypic variability exists for seed diameter, seed thickness and seed plumpness. Association mapping (AM) is a QTL mapping method that can accommodate broader phenotypic variation versus bi-parental populations like LR-18. Association mapping identifies QTLs through linkage disequilibrium. Linkage disequilibrium (LD) is the non-random association of alleles at different loci. If a marker is in LD with a gene locus controlling a trait, they can be statistically associated. Unlike F_2 or RIL populations, the association panels used in AM consist of individuals that are not directly related and have gone through recombination at far greater levels, which can result in a finer mapping resolution of QTLs.

For this study an association panel was assembled, of individuals displaying large amounts of phenotypic variation, to understand and identify SNP marker loci associated with seed size and shape in lentil. The objectives of this study were to 1) determine the levels of population structure and familial relatedness in the panel, 2) determine the level of linkage disequilibrium in cultivated lentil and 3) identify SNP markers associated with seed size, seed shape and DTF.

5.2 Materials and Methods

5.2.1 Plant Material

A total of 143 lines in $1m^2$ plots in RCBD with three replicates were grown at two different locations near Saskatoon, SK (SPG and Preston) in 2011. These lines included four breeding lines and 52 cultivars from the Crop Development Center

(CDC) in Saskatoon SK. The material that came from the CDC is described in Table 5.1. The origin of all the accessions is listed in Appendix 1.

Table 5.1. Market classes of the CDC material used in this study. Many of the landraces do not fit the criteria for the Canadian market classes and as a result were not included in this table.

CDC material	No. of lines
Small green	7
Medium green	6
Large green	11
Extra small red	7
Small red	16
French green	2
Non-standard market class	3

In addition, 91 landraces obtained from the genebanks at ICARDA in Aleppo, Syria and the USDA-ARS in Pullman, Washington were analyzed. All the lines that were used in this panel differed considerably in seed size and shape (Figure 5.1.)

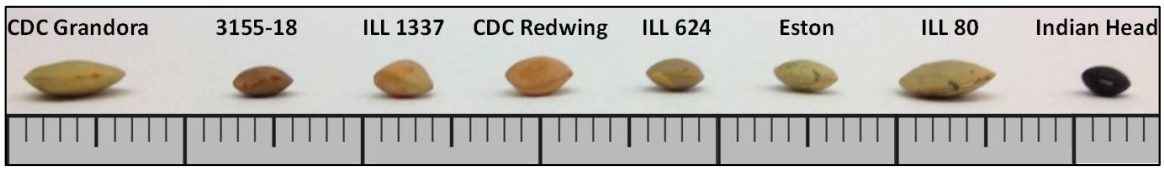

Figure 5.1. Example of differences in seed size and shape among the material grown in the association panel.

5.2.2 Phenotyping

The mature seed samples were measured for seed diameter and seed thickness using round-hole and slotted sieves, respectively, as described in Chapter 4. Seed plumpness was determined by dividing the seed diameter by the seed thickness. The number of days to 50% flowering was recorded for each plot.

5.2.3 Genotyping

Leaf tissue was collected from the field from multiple plants of each genotype. The DNA was extracted using a CTAB extraction method and quantified (Doyle and Doyle 1990). The lentil samples were genotyped with the Lc1536 GoldenGate array as described in Chapter 3. All genotyping information is available through the KnowPulse web portal (http://knowpulse2.usask.ca/portal/).

5.2.4 Linkage Disequilibrium

In total, 451 SNP markers were polymorphic in both the association panel and the RIL population. Only 305 markers were chosen to estimate the linkage disequilibrium (LD) in lentil, because many of the markers mapped very close, or to the same position. The software package Graphical Genotypes (GGT 2.0) (van Berloo 2008) was used to calculate pair-wise r^2 values for markers within each linkage group. Visualization of LD decay was plotted using the program R (R Development Core Team 2011). LD was measured from the combined results of the whole panel as well as with cultivars and landraces separated. Decay of LD over the genetic distance (cM) was calculated by plotting the pair-wise r^2 values, between markers on the same linkage group, over their genetic distance (cM). A second-degree locally weighted scatterplot smoothing (LOESS) line was fitted to each figure. The relationship between LD and genetic distance was evaluated using a method whereby a fixed r^2 value of 0.1 was used a baseline (Robbins et al. 2011). Estimation of LD decay was at the point that the LOESS curve first intercepts the baseline r^2 value.

5.2.5 Phylogenetic Tree Construction

Genetic distance was calculated using Nei's (1972) standard genetic distance measurement using the program SPaGeDi (Hardy and Vekemans 2002). A total of 1000 individual bootstrap replications were performed. The phylogenetic tree was constructed using UPGMA and visualized in the program TreeView (Page 1996).

5.2.6 Population Structure and Kinship Calculations

The program STRUCTURE v2.2 (Pritchard 2000) was used to calculate the number of sub-populations in the panel. The panel was analyzed using the admixture model, with a burn-in time of 50,000. The number of Markov chain Monte Carlo repetitions was set to 50,000. The number of K runs was set from 2-10, with five iterations for each K value. The number of groups that was selected was based on the procedure used by Evanno et al. (2005). The values for each K were submitted to the STRUCTURE harvester website (http://taylor0.biology.ucla.edu./struct_harvest/) which returned the Δ K value

(Earl et al. 2012). The group that had the highest ad hoc statistic Δ K value was selected.

SPAGeDi (Hardy and Vekemans 2002) was used to create a kinship coefficient estimation matrix, with negative values between individuals set to 0.

5.2.7 Association Analysis

The software program TASSEL (Bradbury et al. 2007) was used for the association analysis. To reduce false or spurious associations, population structure (Q) and kinship (K) were calculated first. They were used as covariates in a mixed linear model (MLM) for the associations. A generalized linear model (GLM) was also used where only the Q was used as a covariate. The significance levels were modified using the Bonferroni correction, where each significance value was divided by 982, the number of markers used. Anything below the corrected <0.05 p-value was considered significant.

5.3 Results

5.3.1 Phenotypic Data

For all traits measured there were significant differences amongst the genotypes (Table 5.2). Location was significant only for seed thickness and seed plumpness. The genotype by location interaction was also significant for all traits. Due to the interaction of the traits with the environment, associations were analyzed for each site separately.

Table 5.2. F-values for seed diameter, seed thickness, seed plumpness and DTF.

		F – values			
Effect	df	Diameter	Thickness	Plumpness	DTF
Genotype	132	135.3***	13.7***	83.6***	7.86***
Location	1	13.4ns	84.0*	25.5*	16.58ns
Genotype * Location	132	3.2***	3.7***	3.6***	1.84***
CV%		14.06	5.05	8.55	8.54

*** P≤ 0.001, **P≤ 0.01, *P≤ 0.05, ns not-significant

Broad-sense heritability was calculated for each trait using variance components (Table 5.3). Both seed diameter and seed plumpness showed high

heritability of 0.94 and 0.97, respectively. Seed thickness and flowering time were more moderate with both having a heritability of 0.62.

Table 5.3. Variance components and broad-sense heritability of seed diameter, seed thickness, seed plumpness and DTF.

Variance Component	Diameter	Thickness	Plumpness	DTF
σ^2G	0.37	0.0062	0.0034	2.32
σ^2G * Loc	0.0067	0.0018	0.00012	0.37
σ^2E	0.0075	0.002	0.00012	1.07
σ^2P	0.38	0.01	0.0036	3.76
H^2	0.97	0.62	0.94	0.62

Pearson's correlation coefficients were calculated between all the traits measured (Table 5.4). Seed diameter and seed plumpness showed the highest significant correlation of r=-0.91. The only non-significant correlation was between seed thickness and seed plumpness. Flowering time showed significant but lower correlations with all the seed traits.

Table 5.4. Pearson's correlation coefficients for seed diameter, seed thickness, seed plumpness and DTF.

	Diameter	Thickness	Plumpness	DTF
Diameter	-			
Thickness	0.39***	-		
Plumpness	-0.91***	0.012ns	-	
DTF	0.23***	-0.18***	-0.31***	-

5.3.2 SNP Genotyping

A total of 1049 SNP markers were found to be polymorphic among the entries in the diversity panel. Three genotypes, ILL 4782, ILL 5490 and PI 320953, all had >50% missing data, and were removed from the analysis. The polymorphic markers were compared with the genetic map from the LR-18 population, 451 markers were polymorphic in both the mapping population and the diversity panel. These markers were evenly distributed throughout the linkage map on all seven

linkage groups. For the association analysis, only markers with greater than 10% allele frequency (982) were used for the association analysis (Table 5.5).

Table 5.5. SNPs used in the AM study that mapped in the LR-18 linkage map, and the number of SNPs with >10% allele frequencies.

Comparison with LR-18 linkage map		
Linkage group	Number of mapped markers	markers >10% allele frequency
LG-1	74	71
LG-2	70	70
LG-3	70	61
LG-4	65	62
LG-5	60	55
LG-6	29	27
LG-7	83	76
Unmapped	628	560
Total	1079	982

5.3.3 Linkage Disequilibrium

The LD was calculated for all the lines in the panel and then separately for the cultivars and the landraces. For the whole panel, a total of 45,754 pair-wise comparisons were made. In total, 7,831 or 17% of loci pairs were in significant LD ($p<0.001$). The breeding material and the landraces were also individually examined. The number of significant loci pairs for the breeding material was 6,094 (13%) and 1914 (4%) for the landraces.

All pair-wise comparisons were plotted over the genetic distance (cM) and fitted with a LOESS curve. Differences between the combined panel, the breeding material and landraces were observed (Figure 5.2.). The LD amongst the landraces appeared to decay the fastest <5cM, while the combined panel showed LD decay around 5cM. The LD for the breeding material was the longest at around 20cM.

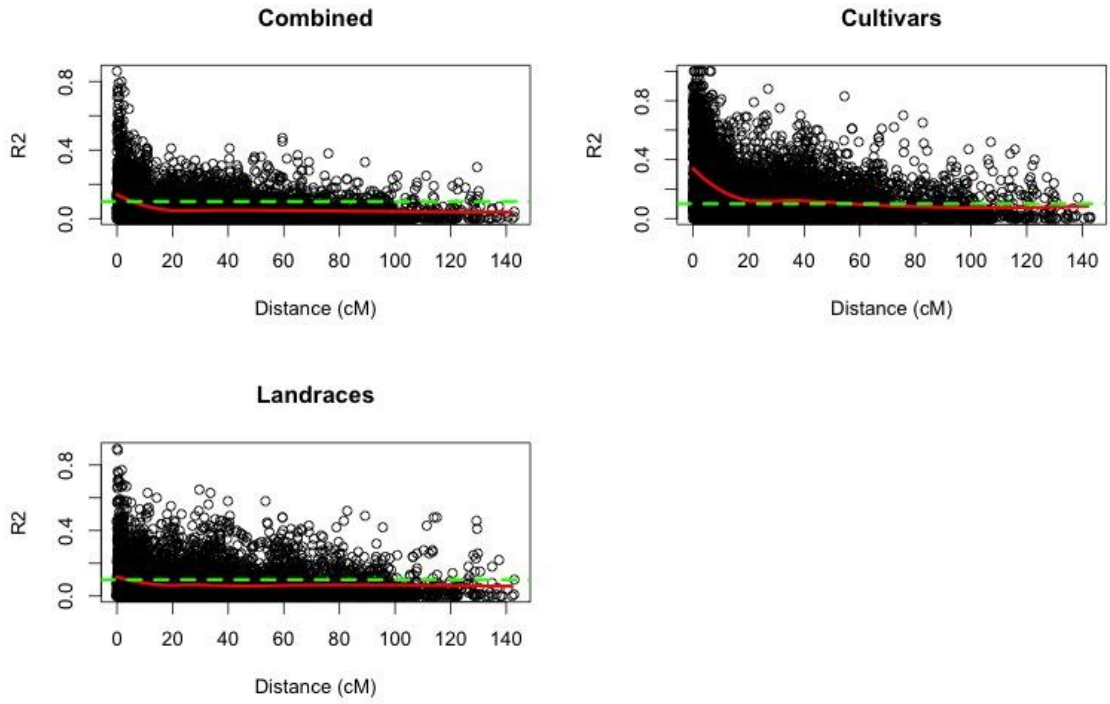

Figure 5.2. Linkage disequilibrium (r^2) plotted against genetic distance (cM) for pair-wise comparisons of markers located throughout the genome. The red lines indicate the second-degree LOESS that was fit for each plot and help represent the rate of LD decay. The green dashed line represents the fixed r^2 value of 0.1. Any value above 0.1 is considered in LD.

5.3.4 Population Structure

The population structure of the panel was calculated using the program STRUCTURE. The highest ad hoc statistic Δ K value was K=4 (Figure 5.3) and corresponded best with the different gene pool origins and breeding history of the lines. Results of each individual and their level of admixture amongst each of the groups are listed in Appendix 2. The two largest groups were groups 1 (red) and 4 (yellow), containing 35% and 30% of the lines, respectively. Group 2 (green) contained 21% while group 3 (blue) was the smallest group containing 14% of the lines. Groups 1 and 3 were mainly from the breeding/elite material, while the majority of the landraces appeared in group 2 and 4. However, none of the groups were completely made up of either the breeding material or the landraces (Figure 5.4.).

The four different groups appeared to be separated by their seed size as well as their history of breeding. Group 1 had a mean seed diameter of 4.23 (mm) while group 3 had a mean seed diameter of 5.01 (mm). The two groups consisting mainly of landraces, group 2 and 4, had a mean seed diameter of 4.89 and 4.01, respectively. However, there were no significant differences among the groups for their seed diameter due to the variablity within each group (Figure 5.5.)

The groups also appeared to reflect some regional differences. Analysis of quantitative morphological traits has revealed that there are four regional groups that lentil accessions come from (Erskine et al. 1989): the Levantine group (Syria, Lebanon, Jordan, Egypt), the northern group (Europe, South America, Iran, Turkey and the former USSR), the Indian group (India, Bangladesh) and the Ethiopian group (Ethiopia). Of the 19 landraces that were in group 1, ten of those were collected in Europe and three were collected in South America, both belonging to the northern group. The remaining three were from the Levantine group of countries. In group 2, the majority of the accessions came from the Levantine group. For group 3, which had only one landrace, the accession came from Russia (northern group) while the rest of the lines were derived from the CDC. Group 4 contained mainly accessions from the Levantine group and from the Indian group. Group 4 also contained accessions from central Asia (Afghanistan, Pakistan). It is unclear as to which regional group those accessions would belong. The accessions that seemed to group with the cultivars and breeding material the most were the northern group of accessions. Adaptation to longer day length or cooler climates may point to the similarities between European germplasm and the North American cultivars.

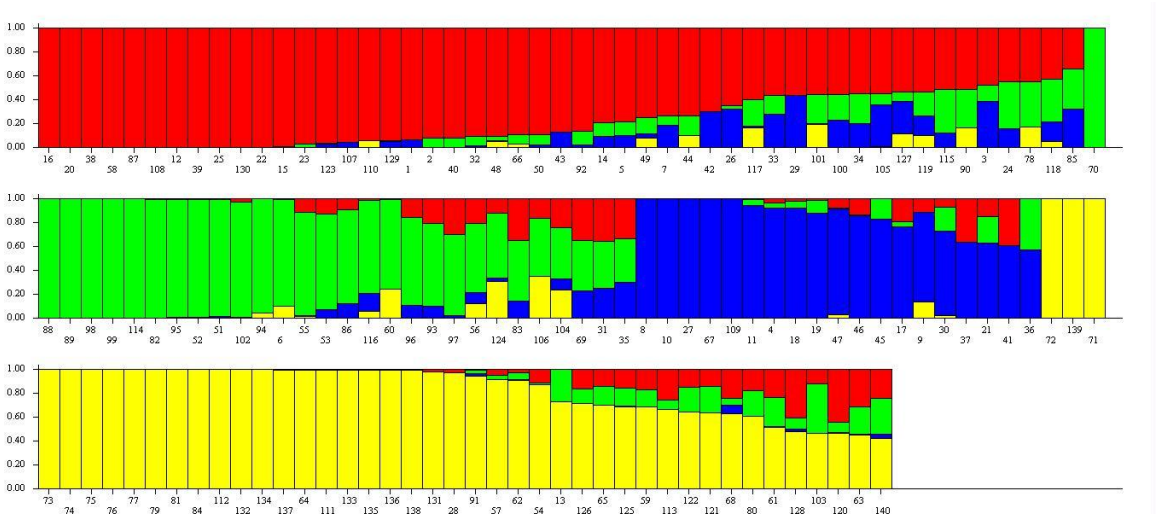

Figure 5.3. Subpopulations and admixtures of 140 lentil lines genotyped and sorted into populations based on STRUCTURE analysis. Each bar represents the individual while the color represents the subpopulation and admixture of each individual.

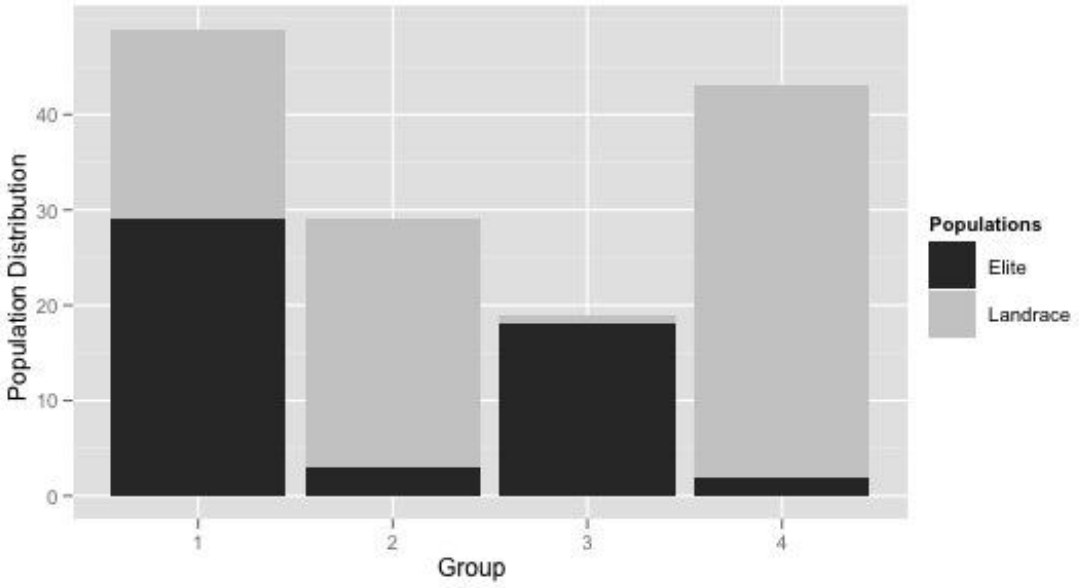

Figure 5.4. Contribution of each ancestral population to the STRUCTURE groups. Group 1 consists of elite breeding lines but does carry a significant amount of landraces. Groups 2 and 4 mainly consist of landraces while group 3 is predominantly elite breeding lines.

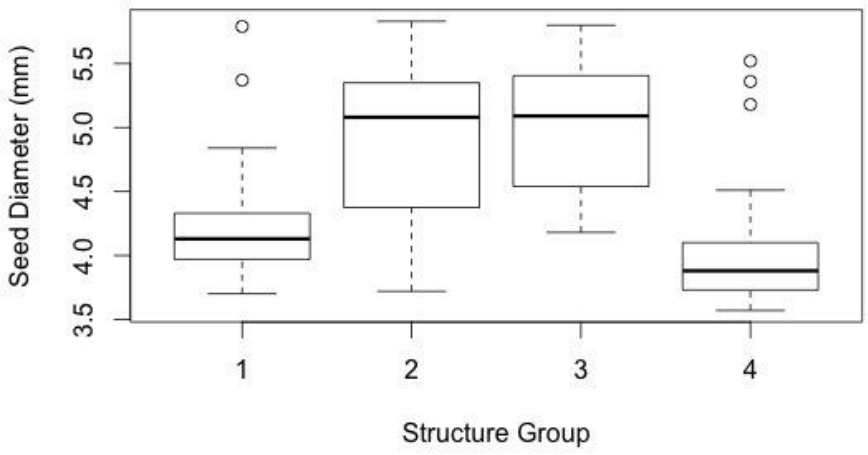

Figure 5.5. Distribution of seed diameter for each of the population groups identified in STRUCTURE.

The UPGMA clustering analysis, which is based on genetic distance, also indicates that the lines are grouped mainly by their breeding history and their seed size (Figure 5.6). There were two major clusters which then split into two additional clusters each. Individuals from population groups 1 and 3 based on STRUCTURE were found mainly in one major cluster, suggesting that there is more genetic similarity between individuals in those two groups. Individuals from groups 2 and 4 typically fell into the second major cluster. Individuals that showed high levels of admixture in STRUCTURE, appeared to form additional clusters in the dendogram (Fig. 5.6 – un-highlighted individuals). There were some lines that deviated from their respective groupings. For example, CDC Redcoat appeared very distant from the remaining cultivars and did not cluster with anything else.

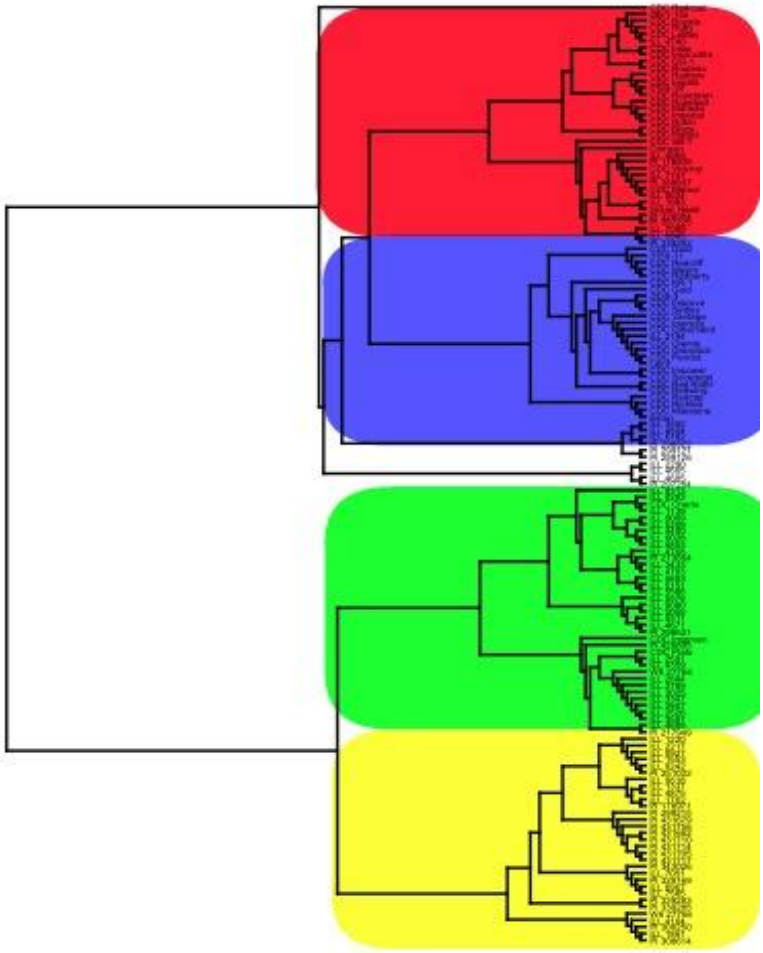

Figure 5.6. UPGMA dendogram of the lentil diversity panel constructed using NEI's (1972) standard genetic distance measurement method. Regions surrounded by different colours correspond to the different sub-groups constructed using STRUCTURE. Un-highlighted clusters were admixtures indicating hybrids between groups.

5.3.5 Association Analysis

Two different models were used in the association analysis: the generalized linear model (GLM) and mixed linear model (MLM). The GLM model takes the population structure (Q) into consideration, while the MLM model uses both the Q and kinship (K). Because there was an interaction between the genotypes and the environment, the two different locations were analyzed separately. For the GLM model 31 different associations were observed between molecular markers and traits observed (Table 5.6). There were no significant associations with DTF. Fifteen of the associations were for seed diameter; nine were for seed plumpness and six for

seed thickness. Five of the markers were significant at both locations. The marker LcC07680p141 was significant at both locations for seed diameter and seed plumpness. Eight markers had been previously mapped to the LR-18 linkage map (Chapter 3), while six markers had mapped in the lentil RIL population LR-139 (Eston x PI 320937) (Alahakoon unpublished). Based on the MLM model, no associations were observed for either location.

The cumulative p-value distributions were plotted with both the GLM and MLM models and with a third model, in which neither population structure nor kinship was accounted for. This method can help identify which models are correcting for false positives more accurately. The model without any population or kinship control skewed towards a greater number of significant associations, which would increase the false positive rate (Figure 5.7). The GLM model shows a greater number of significant associations versus the MLM model. This is consistent with the results observed, where the GLM model had a greater number of significant associations while the MLM model had no significant associations.

Table 5.6. Significant marker associations with corrected p-values for seed diameter, seed thickness and seed plumpness estimated with the GLM model using SNP genotyping data for 140 diverse lentil lines.

		GLM	
Trait	Marker	p-value	Site
Diameter	LcC07680p141	1.15E-08	SPG, Preston
Plumpness	LcC07680p141	6.31E-08	SPG, Preston
Plumpness	LcC05744p200	2.11E-06	Preston
Thickness	LcC22093p71	2.25E-06	Preston
Diameter	LcC05904p141	4.29E-06	SPG, Preston
Diameter	LcC03720p331	4.79E-06	Preston
Diameter	LcC06440p353	5.17E-06	SPG, Preston
Diameter	LcC02075p411	6.50E-06	Preston
Plumpness	LcC03720p331	7.82E-06	Preston
Thickness	LcC17848p320	1.36E-05	Preston
Plumpness	LcC08413p299	2.16E-05	Preston
Plumpness	LcC17429p401	2.21E-05	Preston
Diameter	LcC01215p275	2.59E-05	SPG
Diameter	LcC22735p303	2.60E-05	Preston
Diameter	LcC04319p375	2.87E-05	Preston
Diameter	LcC09496p566	3.00E-05	Preston
Thickness	LcC01398p286	3.20E-05	SPG
Plumpness	LcC05904p141	3.38E-05	SPG
Plumpness	LcC00092p540	3.90E-05	Preston
Plumpness	LcC00632p381	4.32E-05	Preston
Thickness	LcC09496p566	4.98E-05	Preston
Diameter	LcC01869p618	5.02E-05	Preston
Diameter	LcC23126p356	5.56E-05	Preston
Plumpness	LcC23126p356	6.04E-05	Preston
Thickness	LcC00015p324	6.18E-05	Preston
Diameter	LcC00092p540	6.59E-05	Preston
Diameter	LcC08413p299	6.66E-05	Preston
Diameter	LcC07534p343	6.99E-05	Preston
Diameter	LcC06602p283	7.14E-05	SPG, Preston
Thickness	LcC00555p360	7.36E-05	Preston
Thickness	LcC01693p250	9.94E-05	Preston

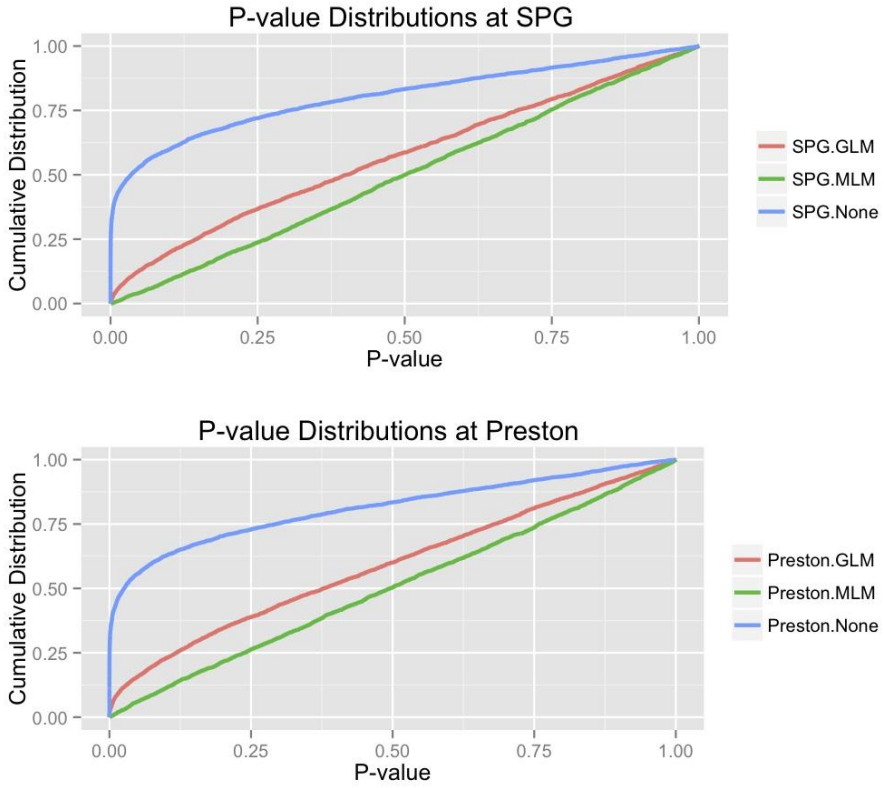

Figure 5.7. Cumulative p-value distributions for the three different association models used for each location (SPG and Preston). A model without any population structure or kinship control (none) is compared to a generalized linear model (GLM) with population control and a mixed linear model with population and kinship control (MLM).

5.4 Discussion

Association mapping was used in this study to identify markers associated with seed morphology and DTF in a diverse panel of breeding lines, cultivars and landraces.

Linkage disequilibrium decay is used to estimate the number of markers that are needed to saturate the genome for association mapping. The LD for the whole association panel was calculated, as well as for the landraces and breeding lines/cultivars separately. Domestication can lead to gene pools with different allele frequencies, which can cause certain allele combinations to change, leading to more extensive LD (Hamblin et al. 2010). This explains why LD in domesticated populations is generally higher than in wild populations. However, large differences can still exist within domesticated populations. For example elite breeding lines

have gone through many bottlenecks and would still exhibit much higher levels of LD compared to landraces. In corn, LD decay has been noted to exist as high as 100kb in elite inbred lines (Ching et al. 2002). In contrast, LD decayed within 1kb (Tenaillon et al. 2001) in diverse landraces. In this study, the genome-wide LD was higher for the breeding material and lower for the landrace material, which supports that assumption. This suggests that landraces of lentil could be used in further association studies, leading to marker associations that would be closer to the polymorphism causing the phenotypic variation. Nonetheless, the LD decay for the landrace material that was measured could still be considered high suggesting it would perhaps be better to go further into wild material to identify closer associations. Self-pollinated crops like lentil are expected to have higher levels of LD when compared to outcrossing crops like corn. LD has been noted to extend beyond 10cM in barley cultivars and even 50-100cM in local populations of Arabidopsis (Kraakman et al. 2004; Nordborg et al. 2002). In a panel of elite wheat cultivars, LD was reported to decay at a distance of 5cM (Somers et al. 2007). Insight into why the LD levels appeared high may come from the genetic map that was used to determine the distances between markers. The linkage map that was used in the calculation of LD over distance was developed in only one population, LR-18. This map does not represent the entire genetic diversity in *L. culinaris* ssp. *culinaris* because its calculation is based on only the RILs derived from the two parents of the population. This linkage map also contains highly clustered regions separated by large gaps (>10cM). Whereas in other studies that have calculated LD over distance, the maps used were usually consensus maps that were much more condensed with markers and represented greater genetic diversity of the species (Somers et al. 2007; Soto-Cerda and Cloutier 2012). However, for lentil there is currently no consensus map available. In addition, LD was also calculated using genetic distance instead of the physical distance. Ersoz et al. (2007) notes that LD should be calculated using the LD decay from a physical distance on a number of loci, instead of a whole genome. For example, some loci which may be linked to domestication or other highly selected traits, could be under high levels of LD, which could inflate the LD estimate for the rest of the genome. Nonetheless, nearly 50% of the loci showed significant levels of

LD. This should mean that there are sufficient levels of LD in lentil for effective association mapping to occur with the current SNP marker density.

Population structure revealed that four sub-groups were present in the panel (Figure 3.). Each of the groups appeared to represent their breeding history with groups from elite breeding material and landraces forming different sub-groups. All groups, however, were not completely composed of breeding lines or landraces. For example, of the 49 lines in group 1, 20 were landraces (Figure 5.4.). Only three cultivars were placed into group 2. They were CDC Cherie, CDC Redberry and CDC Redcoat. All three belong in the small red cotyledon market class and share similar pedigrees. An explanation for why those three cultivars were grouped into group 2 remains unknown because in each of their pedigrees no crosses with any of the other accessions in group 2 were made. In group 3 the only landrace was accession ILL 2194. A possible reason why this accession grouped amongst cultivars is that it belongs to the northern group of accessions, which share similar seed morphology, such as large seed diameter with a green seed coat. Group 4 contained two cultivars CDC Imigreen and CDC Plato. When the pedigrees of these two cultivars were analyzed there was no obvious explanation as to why they were in group 4. UPGMA tree-based analysis, in large part, did support the sub-groups that were formed in STRUCTURE. The clusters were similar to the population sub-structures with each of the four main sub-groups forming nearly the same clusters. However, there were some lines that did not group in the dendogram according to population structure estimates. Liu et al. (2008) also measured the population structure of lentil using SSR markers. They identified 8 sub-groups using 440 accessions from the Chinese National Gene Bank. In this case, the groups were separated mainly on their geographic origins. For example, germplasm from India and North America formed independent groups. Genetic diversity studies using UPGMA, determined by the genetic distance between lines, have also revealed that the geographic origin is a primary variable in explaining the variation amongst lentil germplasm (Alabboud et al. 2009). The results of this study also confirmed that geographic origins influence population structure in lentil. Groups 1 and 3 were mainly made up of North American adapted cultivars and the northern group of accessions, while groups 2

and 4 were composed mainly of accessions from the Levantine, Indian and Ethiopian groups. However, in previous genetic diversity and population structure studies the authors did not expand on any possible phenotype characteristics that could further explain each group. The morphological characteristic that has traditionally distinguished cultivated lentil is seed size. Barulina (1930) was the first to distinguish lentil germplasm based on seed size, describing large seeded lentils as macrosperma and small seeded types as microsperma. These two types of lentil have commonly been referred as two different gene pools within the *Lens culinaris* ssp. *culinaris* species that were formed during domestication. The results from this study at first appeared to indicate that seed size distinguishes sub-groups within L. *culinaris* ssp. *culinaris.* However, when plotted there were no significant differences amongst the groups for seed diameter. These results are similar to the findings of Alo et al. (2011). This study measured the population structure of the *Lens* genus. The two *L. culinaris* ssp. *culinaris* sub-groups that were analyzed differed in seed weight, with larger and smaller sized seeds forming separate groups. There was, however, overlap in the distributions of the two groups and according to a t-test they were non-significant. A chi square test of the frequency distribution for the two groups, however, did show significant difference. Sharma et al. (1995) also observed clustering for seed size within the L. *culinaris* ssp. *culinaris* species using RAPD markers. Studies using RAPD and ISSR markers, found no clear clustering pattern between micro and macrosperma lentils (Abo-elwafa et al. 1995; Sonnante and Pignone 2001). However, these studies used limited number of markers and small population sub-sets.

After the markers were corrected, in order to reduce the number of false positives, only the GLM method showed significant marker associations; whereas the MLM model did not. The difference between the two models is that the MLM is more stringent by containing a correction for kinship (K) while the GLM does not. The cumulative P-value distributions (Figure 5.7) showed that the GLM was skewed closer in the direction where no model was used. This shows that the MLM model has as more significant fit and that there is a greater chance of spurious associations in the GLM model. However, when comparing the significant markers in the GLM

with the other significant markers in the bi-parental QTL study it shows that those markers map to the same position or near the same QTLs. Other studies have noted differences in the number of significant associations between the two models. Neumann et al. (2011) noted that differences between the two models appeared to be trait dependent. For example, Yu et al. (2005) compared many different models, including the Q and Q+K models, for flowering time, ear height and ear diameter in corn. They found that the Q+K model showed the highest power for flowering time and ear height while for ear diameter the K model showed the highest power. The power of the Q model was also high for flowering time and ear height, but low for ear diameter. This suggests that neither model is ideal for every trait and every species. However for lentil, where the degree of population structure is still largely unknown, the Q+K model is more flexible by having the ability to account for both population and family based structures.

A total of 31 markers were significant in the GLM model after the p-value corrections. The most significant marker, LcC07680p141, was associated with seed diameter and seed plumpness. This marker did not map in the LR-18 linkage map. However, after comparing the contig sequence of the marker with *Medicago truncatula* using BLAST, a homolog of this marker was located on chromosome 5 on the *Medicago* physical map. QTLs for seed weight in *Medicago* have also been associated on that chromosome (D'Erfurth et al. 2012). The marker LcC05904p141, significant only for seed diameter, mapped ~600kbp away from the Pea *Rug3* gene homolog in *Medicago* (Harrison et al. 1998). This gene helps regulate starch synthesis in developing seeds with mutants having small shriveled seeds. Markers associated with seed diameter and seed plumpness mapped ~200kbp from a sucrose transporting gene (*SUT1*) on chromosome 4 in Medicago. Homologs of this gene have been known to control seed development in pea and faba bean (Borisjuk et al. 2008).

One of the primary advantages of AM is the association of loci that may not have been significantly associated with a trait in linkage mapping. This is because large numbers of markers may not be polymorphic between the parents of a bi-population. For this study, it was anticipated that there would be a greater number

of loci discovered that were associated with only seed plumpness. A total of six markers shared significant associations with seed diameter and seed plumpness. However, the marker LcC05744p200 showed high significance for seed plumpness but none for seed diameter. The contig sequence of this marker mapped to chromosome 5 of Medicago. No other contig sequences for significant seed diameter markers mapped nearby. There is potential for this marker to be used to select for seed plumpness without unintentionally selecting for seed diameter. An additional marker, LcC09496p566, was significant for both seed diameter and seed thickness at the Preston location.

In this study, no significant DTF QTLs were identified by either the GLM or MLM model. In other legumes that are close relatives of lentil, such as pea, there is abundance flowering time related loci (>20) (Weller et al. 2009). This could provide some detail why there were no significant associations for DTF. One of the limitations of association mapping is that rare or low frequency alleles often do not have enough statistical significance to be detected. In a diverse collection, such as the one used for this study, there could be many loci with many alleles that are controlling flowering time, which could lower the detection power. Alternatively, as the data is from only one growing season, there could simply be insufficient variability. This could also be contributing to why there were no associations for DTF.

Chapter 6

General Discussion

6.1 Conclusions and Future Work

The overall hypothesis of the study was that SNP markers could be associated with seed size and shape QTLs using linkage and association mapping techniques. This would result in markers that could be used for MAS to select for seed size and shape in developing cultivars.

Lentil, until recently, was considered an orphan crop with regards to genomic research. However this changed with the development of an Illumina 1536 GoldenGate SNP assay by Sharpe et al. (2013). This was the first study in lentil to implement the Illumina 1536 GoldenGate assay. This assay proved to be a highly functional tool that could be used to genotype hundreds of individuals, using thousands of markers, in a short time. In addition to the RIL population LR-18 being genotyped, an association panel was also genotyped. The genetic map constructed from the LR-18 population was the first intraspecific linkage map in lentil to form 7 linkage groups that most likely correspond to the 7 chromosomes of lentil. Also, when compared to other maps, the intraspecific map that was developed resulted in a higher concentration of markers throughout the map. The map also showed regions that were separated by >10cM. This could explain the high LD decay calculated for the association panel (Chapter 5). A consensus map for lentil was proposed to help close those gaps and provide markers that represent more of the lentil genome than just the bi-parental population used to construct the current linkage map. Separate populations, LR-139 (Eston x PI 320937) and LR-03 (ILL 1704 x ILL 7537), have been genotyped with the same assay with the hopes that the markers that were not polymorphic in LR-18 will map to those populations. The three maps, LR-18, LR-139, and LR-03 could then be combined to further increase the resolution of the map and shorten many of the gaps in the current map.

A method, whereby each seed sample was placed on a flat-bed shaker and passed through screens, was used to quantify the seed diameter, thickness and plumpness. In both the mapping population LR-18 and the association panel there

were significant differences amongst all genotypes for seed diameter, thickness and plumpness. This suggests that the screening method was capable of detecting enough of the variation. However, seed diameter and seed plumpness were highly correlated, in both the mapping population and the association panel, and as a result shared many of the same QTLs. If there are separate loci controlling both those traits then high correlations between the two make it more difficult to associate those loci with markers. Seed diameter and seed plumpness have a natural correlation, but it is suspected that since seed plumpness was determined by dividing the values of seed thickness by seed diameter, there may have been an artificial inflation in the correlation between those traits. There are an increasing number of phenotyping software options such as TomatoAnalyzer (Rodriguez et al. 2010) and SmartGrain (Tanabata et al. 2012), that could be used to phenotype seed plumpness in lentil. In addition, methods developed at the Canadian Grain Commission have been developed to estimate seed plumpness using digital images (Shanin et al. 2006). At least in the LR-18 population, seed plumpness may not have been segregating. It was proposed that a population in which the two parents have the same seed diameter, but differ in their seed plumpness would result in the population segregating solely for seed plumpness. This would allow seed plumpness QTLs to be more accurately mapped. Those methods could be implemented to increase the accuracy of seed plumpness estimation and potentially decrease it's correlation with seed diameter. This would result in the identification of QTLs that independently control each of those traits.

Heritability estimates for all the traits that were measured were similar between the LR-18 population and the AM panel. In the LR-18 population heritability estimates of seed diameter, seed thickness, seed plumpness, and DTF were 0.92, 0.60, 0.94 and 0.45, respectively. While in the association panel they were: 0.97, 0.62, 0.94 and 0.62 for seed diameter, seed thickness, seed plumpness and DTF, respectively. The heritability estimates for seed diameter and seed plumpness were very high. One of most useful properties of MAS is efficiently selecting traits that have a low heritability. For traits with high heritability estimates, like seed diameter and seed plumpness, MAS may not be as useful.

However, the lentil breeding program at the CDC often makes three-way crosses when developing cultivars. If male parents in the final cross are heterozygous this would result in heterogametic progeny (Singh, 1994). It has been proposed that gamete selection of those F_1 progeny could help enrich the populations to improve the material that would be selected in the field. The markers that were associated with the highly heritable traits could still then be used for MAS.

Estimation of linkage disequilibrium (LD) is an important calculation that is needed prior to association mapping to determine the number of markers required. Markers that were located on the LR-18 map were used to estimate the LD over genetic distance in the association panel. Results showed that nearly 20% of all pair-wise LD calculations between markers in the association panel were significant. Genome wide LD decay was estimated to be ~5cM, which would be enough to implement genome-wide AM. However, the extent of LD suggests that the markers that were associated with traits are unlikely to be the causal polymorphism controlling the trait. LD for the landraces and breeding material was also calculated separately. The genetic distance in which LD decayed was lower for the landraces, suggesting that landraces could be used in further AM studies to locate markers closer to the causal polymorphism.

DTF was measured to determine if any QTLs were co-located with the other seed size and shape QTLs. In Chapter 4, a DTF QTL was co-located with both seed diameter and seed plumpness QTLs at the *Yc* locus in the LR-18 population. If that marker was used to select for seed diameter and plumpness, indirect selection for DTF would result. There were two more QTLs, on linkage groups 2 and 7, associated with seed diameter and plumpness that did not co-located with QTLs for DTF. In Chapter 5, no significant marker associations with DTF were detected, suggesting that DTF and seed diameter and plumpness QTLs were not co-located. In the AM panel, further evidence showed that seed diameter and plumpness were also segregating with the *Yc* locus, leading us to speculate that the previous DTF QTL linkage to the *Yc* locus could also be broken. Nonetheless, this study confirmed, at least in some populations, that DTF can co-segregate with other seed size and shape QTLs. There were though, other QTLs for seed size and shape that were not linked to

DTF loci. Markers linked to those QTLs would be suitable for MAS and would not indirectly select for DTF.

By using both the association and linkage mapping strategies we were able to confirm several genomic regions controlling each of the traits and identify potential false positives. A total of eight markers identified in the AM study (Chapter 5) were located also identified in the LR-18 linkage map (Table 6.1). Six additional markers also mapped in the LR-139 population. However, there have been no attempts to map seed size or shape QTLs in this population. Five of the markers that were significant in the AM study (Chapter 5), were also significant for QTLs in the population LR-18 (Chapter 4), confirming their association with those traits. For seed diameter there were three QTLs detected on linkage groups 1, 2 and 7 in the linkage mapping study. In the AM study there were a total of 15 significant markers for seed diameter. Of those markers that were significant in the AM study (Chapter 5), three mapped in LR-18. Furthermore, those markers only mapped to linkage groups 1 and 7. When compared to the QTLs for seed diameter in LR-18, the markers that were significant in the AM study were closely located. For example, marker LcC05904p141 mapped only 3 cM away from the *Yc* locus, which was significantly associated for seed diameter, seed plumpness and DTF in the LR-18 population. However, there were significant markers that were in the LR-18 map but showed no association with QTLs shown in the AM study. The marker LcC03720p331, which was significant for both seed diameter and seed plumpness, mapped to linkage group 7 at a locus nearly 6cM away from the QTLs for seed diameter and seed plumpness in LR-18. The most consistent QTL for seed thickness in LR-18 appeared on linkage group 7. None of the markers associated with seed thickness in the AM study appeared on linkage group 7, except the marker LcC09496p566. This marker mapped just over 2cM away from the QTL in the linkage map. Two markers, LcC00555p360 and LcC01693p250, mapped to linkage group 3. However, no QTLs related to seed thickness mapped in the linkage mapping study (Chapter 4). Since those markers did not show any association in Chapter 4, the likelihood that they are false positives is higher. Those loci controlling seed thickness could have also been monomorphic in LR-18 and did not segregate.

Table 6.1. Significant markers in Chapter 5 and their positions on the LR-18 and LR-139 linkage maps.

Trait	Marker	LR-139 map	LR-18 map
Diameter	LcC07680p141	unmapped	unmapped
Plumpness	LcC07680p141	unmapped	unmapped
Plumpness	LcC05744p200	unmapped	LG-1
Thickness	LcC22093p71	unmapped	unmapped
Diameter	LcC05904p141	LG-1	LG-1
Diameter	LcC03720p331	unmapped	LG-7
Diameter	LcC06440p353	LG-7	unmapped
Diameter	LcC02075p411	LG-7	unmapped
Plumpness	LcC03720p331	unmapped	LG-7
Thickness	LcC17848p320	unmapped	unmapped
Plumpness	LcC08413p299	unmapped	unmapped
Plumpness	LcC17429p401	unmapped	unmapped
Diameter	LcC22735p303	unmapped	unmapped
Diameter	LcC04319p375	unmapped	unmapped
Diameter	LcC09496p566	unmapped	LG-7
Plumpness	LcC00092p540	unmapped	unmapped
Plumpness	LcC00632p381	unmapped	unmapped
Thickness	LcC09496p566	unmapped	LG-7
Diameter	LcC01869p618	unmapped	unmapped
Diameter	LcC23126p356	unmapped	unmapped
Plumpness	LcC23126p356	unmapped	unmapped
Thickness	LcC00015p324	unmapped	unmapped
Diameter	LcC00092p540	unmapped	unmapped
Diameter	LcC08413p299	unmapped	unmapped
Diameter	LcC07534p343	unmapped	unmapped
Diameter	LcC06602p283	LG-1	unmapped
Thickness	LcC00555p360	LG-3	LG-3
Thickness	LcC01693p250	LG-3	LG-3

Seed plumpness and seed diameter QTLs were located at many of the same loci in both studies. It would be difficult to select lines, using those markers, for seed plumpness that would be independent of seed diameter. In the AM study, there were two significant markers, LcC05744p200 and LcC17429p401, associated with seed plumpness that showed no association with seed diameter. However, the marker LcC05744p200 did map on linkage group 1 less than 5cM away from LcC05904p141, which was associated with seed diameter. If the calculation of LD decay (~ 5cM) is correct, then those two markers would be in LD. This would still result in the indirect selection for seed diameter when selecting seed plumpness

using that marker. The other marker, LcC17429p401, did not map in the LR-18 or LR-139 linkage map, therefore it's position and its linkage status with other markers is unknown. Further studies will be needed to determine if this marker could be of value.

Each SNP marker carries a contig sequence from which it was developed. Due to the extent of LD in lentil, it is unlikely that those contigs could be proposed as sequences for candidate genes of those traits. However, by using the BLAST program we could identify homologous sequences in model crops that map near candidate genes for other seed size traits. For example, D'Erfurth et al. (2012) noted that a subtilase gene (*SBT1.1*) located on chromosome 5 of *Medicago truncatula* was likely controlling a seed weight QTL. When the contig sequence of LcC00890p1387, which was linked to multiple QTLs, was BLAST against the *Medicago* physical map both loci mapped near one another. This could lead to the identification of candidate genes and further increase our understanding of the genetics of seed diameter, seed thickness, seed plumpness and DTF in lentil. Further genomic information will become available when the lentil genome is sequenced, which is currently underway (K. Bett, pers. comm.). A more informative lentil physical map could result in a greater understanding of the genomic regions controlling traits. For example, the genomic region surrounding the cotyledon colour locus *Yc*, which is linked to multiple QTLs for seed diameter, seed plumpness and DTF in the LR-18 population, could be further examined to determine which specific sequences are controlling those traits.

The results from this study have confirmed that SNP markers can increase the density of the lentil genetic map and further increase the resolution of QTLs that are mapped. It has also increased our understanding of the lentil genome, the future potential of association mapping and the inheritance of seed size and shape in lentil.

References

Abbo S, Ladizinsky G, and Weeden NF (1991) Genetic-analysis and linkage study of seed weight in lentil. Euphytica 58: 259-266

Abbo S, Molina C, Jungmann R, Grusak MA, Berkovitch Z, Reifen R, Kahl G, Winter P, Reifen R (2005) Quantitative trait loci governing carotenoid concentration and weight in seeds of chickpea (*Cicer arietinum* L.). Theor Appl Genet 111: 185-195

Abo-Elwafa A, Murai K, Shimada T (1995) Intra-and inter-specific variations in *Lens* revealed by RAPD markers. Theor Appl Genet 90: 335-340

Alabboud I, Szilagye L, Roman GV (2009) Assessment of genetic diversity in lentil (*Lens culinaris* Medik.) as revealed by RAPD markers. Scientific Papers Series A 2:439-444

Alo F, Furman BJ, Akhunov E, Dvorak J, Gepts P (2011) Leveraging genomic resources of model species for the assessment of diversity and phylogeny in wild and domesticated lentil. J Hered 102: 315-329

Arumuganathan K, and Earle ED (1991) Nuclear DNA content of some important plant species. Plant Molecular Biology Reporter 9: 208-218

Babayeva S, Akparov Z, Abbasov M, Mammadov A, Zaifizadeh M, Street K (2009) Diversity analysis of Central Asia and Caucasian lentil (*Lens culinaris* Medik.) germplasm using SSR fingerprinting. Genet Res Crop Evol 56: 293-298

Barulina H (1930) Lentils of the USSR and other countries (English summary). Bulletin of Applied Botany, Genetics and Plant Breeding 40: 265-304

Batley J, Edwards D (2007) SNP applications in plants. In: Oraguzie NC, Rikkerink EHA, Gardiner SE, De Silva HN (ed) Association mapping in plants, Springer, New York, 95-102

Bates, DM (2010) lme4: Mixed-effects modeling with R. Accessed http://lme4. r-forge. r-project. org/book, November 2011

Bayaa B, Erskine W, Hamdi A (1995) Evaluation of a wild lentil collection for resistance to vascular wilt. Genet Res Crop Evol 42: 231-235

Borisjuk L, Rolletschek H, Radchuk R, Weschke W, Wobus U, Weber H (2008) Seed development and differentiation: a role for metabolic regulation. Plant Biology 6: 375-386

Brachi B, Faure N, Horton M, Flahauw E, Vazquez A, Nordborg M, Gergelson J, Cuguen J, Roux F (2010) Linkage and association mapping *of Arabidopsis thaliana* flowering time in nature. PLoS Genetics 6: e1000940

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23: 2633-2635

Branca A, Paape TD, Zhou P, Briskine R, Farmer AD, Mudge J, et al. (2011) Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*. Proc Nat Acad Sci 108: E864-E870

Breseghello F, Sorrells ME (2006) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. Genetics 172: 1165-1177

Breseghello F, Sorrells ME (2007) QTL analysis of kernel size and shape in two hexaploid wheat mapping populations. Field Crop Resear 101: 172-179

Chagné D, Crowhurst RN, Troggio M, Davey MW, Gilmore B, Lawley C, Vanderzande S, Hellens RP, Kumar S, Cestaro A, Velasco R, Main D, Rees JD, Iezzoni A, Mockler T, Wilhelm L, Van de Weg E, Gardiner S.E, Bassil N, Peace C (2012) Genome-Wide SNP detection, validation, and development of an 8K SNP array for apple. PLOS One 7: e31745

Chant, SR (2004) Imidazolinone tolerance in lentil (*Lens culinaris* Medik.). Dissertation, University of Saskatchewan, Saskatoon, SK

Ching ADA, Caldwell K S, Jung M, Dolan M, Smith OS, Tingey S, et al. (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genetics 3: 19

Close TJ, Bhat PR, Lonardi S, Wu Y, Rostoks N, Ramsay L, Druka A, Stein N, Svensson JT, Wanamaker S, Bozdag S, Roose ML, Moscou MJ, Chao S, Varshney RK, Szucs P, Sato K, Hayes PM, Matthews DE, Kleinhofs A, Muehlbauer GJ, DeYoung J, Marshall DF, Madishetty K, Fenton RD, Condamine P, Graner A, Waugh R (2009) Development and implementation of high-throughput SNP genotyping in barley. BMC Genomics 10:582

Cober ER, Fregeau-Reid JA, Voldeng HD (1997) Heritability of seed shape and seed size in soybean. Crop Sci 37: 1767-1769

Cobos MJ, Rubio J, Fernandez-Romero MD, Garza R, Moreno MT, Millan T, Gil J (2007) Genetic analysis of seed size, yield and days to flowering in a chickpea

recombinant inbred line population derived from a Kabuli x Desi cross. Ann App Bio 151: 33-42

D'Erfurth I, Signor C, Aubert G, Sanchez M, Vernoud V, Darchy B et al. (2012) A role for an endosperm-localized subtilase in the control of seed size in legumes. New Phyto 196: 738-751

Diaz A, Fergany M, Formisano G, Ziarsolo P, Blanca J, Fei Z, et al. (2011). A consensus linkage map for molecular markers and Quantitative Trait Loci associated with economically important traits in melon (*Cucumis melo* L.). BMC Plant Biology 11: 111

Domoney C, Duc G, Ellis TH, Ferrándiz C, Firnhaber C, Gallardo K, et al (2006) Genetic and genomic analysis of legume flowers and seeds. Curr Opin Pl Bio 9: 133-141

Doyle JJ, Doyle JL (1990) Isolation of plant DNA from fresh tissue. Focus 12: 13-15

Duran Y, Fratini R, Garcia P, Perez de la Vega M (2004) An intersubspecific genetic map of *Lens*. Theor Appl Genet 108:1265-1273

Earl DA (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Res 4: 359-361

Environment Canada (2012) http:// www.weatheroffice.gc.ca/canada_e.html. Accessed July, 2012

Eujayl I, Baum M, Powell W, Erskine W, Pehu E (1998) A genetic linkage map of lentil (*Lens* sp.) based on RAPD and AFLP markers using recombinant inbred lines. Theor Appl Genet 97: 83-89

Erskine W, Williams PC, Nakkoul H (1985) Genetic and environmental variation in the seed size protein, yield and cooking qualities of lentils. Field Crops Res 12: 153-161

Erskine W, Adham Y, Holly L (1989) Geographic distribution of variation in quantitative traits in a world lentil collection. Euphytica 43: 97-103

Erskine W, Williams PC, Nakkoul H (1991) Splitting and dehulling lentil (*Lens culinaris*): Effects of seed size and different pretreatments. J Sci Food Ag 57: 77-84

Erskine W (1996) Seed-size effects on lentil (*Lens culinaris)* yield potential and adaptation to temperature and rainfall in West Asia. J Ag Sci 126: 335-341

Erskine W, Sarker A (2004) Lentil. In: Corke H, Walker CE (ed) Encyclopedia of grain sciences. Elsevier, London, UK, pp 142–150

Ersoz E S, Yu J, Buckler E S (2007) Applications of linkage disequilibrium and association mapping in crop plants. In: Varshney RK, and Tuberosa R (ed) Genomics-assisted crop improvement. Springer. Netherlands, pp 97-119

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14: 2611-2620

Fan JB, Chee MS, Gunderson KL (2006) Highly parallel genomic assays. Nature Reviews Genetics 7: 632-644

FAOSTAT (2010) http://faostat.fao.org/. Accessed December, 2010

Flandez-Galvez H, Ford R, Pang ECK, Taylor PWJ (2003) An intraspecific linkage map of the chickpea (*Cicer arietinum* L.) genome based on sequence tagged microsatellite site and resistance gene analog markers. Theor Appl Genet 106: 1447-1456

Ford R, Pang ECK, Taylor PWJ (1999) Genetics of resistance to ascochyta blight (*Ascochyta lentis*) of lentil and the identification of closely linked RAPD markers. Theor Appl Genet 98: 93-98

Ford R, Rubeena RJ, Materne M, Taylor PWJ (2007) Lentil. In: Kole C (ed) Genome Mapping and Molecular Breeding in Plants. Springer-Verlag Berlin Heidelberg, pp 91-108

Ford R, Mustafa B, Baum M, Rajesh PN (2009) Advances in molecular research. In: Erskine W, Muehlbauer FJ, Sarker A, Sharma B (ed) The lentil: botany, production and uses. CABI publishing, Cambridge, MA, pp 155-171

Fratini R, Duran Y, Garcia P, Perez de la Vega M (2007) Identification of quantitative trait loci (QTL) for plant structure, growth habit and yield in lentil. Spanish J Ag Res 5: 348-356

Gegas VC, Nazari A, Griffiths S, Simmonds J, Fish L, Orford S, Sayers L, Doonan JH, Snape JW (2010) A genetic framework for grain size and shape variation in wheat. Plant Cell 22: 1046-1056

Genchev D (2006) Genetic Control of Seed Shape of the Common Bean (*Phaseolus vulgaris* L.). Annual Report of the Bean Improvement Cooperative 49: 169-170

Global Crop Diversity Trust (2012) Global strategy for ex-situ conservation of lentil (*Lens* Miller). http://www.croptrust.org/documents/cropstrategies/lentil.pdf. Accessed December 2012

Gupta PK, Rustgi S, Kulwal PL (2005) Linkage disequilibrium and association studies in higher plants: present status and future prospects. Pl Mol Bio 57: 461-485

Gupta PK, Rustgi S, Kumar N (2006) Genetic and molecular basis of grain size and grain number and its relevance to grain productivity in higher plants. Genome 49: 565-571

Gupta PK, Rustgi S, Mir RR (2008) Array-based high-throughput DNA markers for crop improvement. Heredity 101: 5-18

Gupta D, Taylor PW, Inder P, Phan HT, Ellwood SR, Mathur PN, et al (2012) Integration of EST-SSR markers of *Medicago truncatula* into intraspecific linkage map of lentil and identification of QTL conferring resistance to ascochyta blight at seedling and pod stages. Mol Breed 30: 429-439

Hamdi A, Erskine W, Gates P (1991) Relationships among economic characters in lentil. Euphytica 57: 109-116

Hamblin MT, Close TJ, Bhat PR, Chao S, et al (2010) Population structure and linkage disequilibrium in US barley germplasm: Implications for association mapping. Crop Sci 50: 556-566

Hamwieh A, Udupa S M, Sarker A, Jung C, Baum M (2009) Development of new microsatellite markers and their application in the analysis of genetic diversity in lentils. Breed Sci 59: 77-86

Hancock JF (2004) Plant Evolution and the Origin of Crop Species, 2nd edn. CABI, Wallingford, UK

Hardy OJ, Vekemans X (2002) SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol Ecol Notes 2: 618-620

Harrison CJ, Hedley CL, Wang TL (1998) Evidence that the rug3 locus of pea *(Pisum sativum* L.) encodes plastidial phosphoglucomutase confirms that the imported substrate for starch synthesis in pea amyloplasts is glucose-6-phosphate. Plant J 13: 753-762

Havey MJ, Muehlbauer, FJ (1989) Linkages between restriction fragment length, isozyme, and morphological markers in lentil. Theor Appl Genet 77: 395-401

He C, Tian Y, Saedler R, Efremova N, Riss S, Khan MR, Yephremov A, Saedler H (2010) The MADS-domain protein MPF1 of *Physalis floridana* controls plant architecture, seed development and flowering time. Planta 231:767-777

Holland JB (2007) Genetic architecture of complex traits in plants. Current Opinion in Plant Biology 10:156-161

Hossain S, Ford R, McNeil D, Pittock C, and Panozzo JF (2010) Development of a selection tool for seed shape and QTL analysis of seed shape with other morphological traits for selective breeding in chickpea (*Cicer arietinum* L.). Aus J Crop Sci 4: 278-288

Hovav R, Upadhyaya KC, Beharav A, Abbo S (2003) Major flowering time gene and polygene effects on chickpea seed weight. Plant Breeding 122: 539-541

Jofuku KD, Den Boer BG, Van Montagu M, Okamuro J K (1994) Control of Arabidopsis flower and seed development by the homeotic gene APETALA2. Pl Cell Online 6: 1211-1225

Jones N, Ougham H, Thomas H (1997) Markers and mapping: we are all geneticists now. New Phytol 137: 165-177

Kahraman A, Kusmenoglub I, Aydinc N, Aydogan A, Erskine W, Muehlbauer FJ (2004) QTL mapping of winter hardiness genes in lentil. Crop Sci 44: 13-22

King GJ (2002) Through a genome, darkly: comparative analysis of plant chromosomal DNA. Plant Molecular Biology 48: 5-20

Kraakman AT, Niks RE, Van den Berg PM, Stam P, Van Eeuwijk FA (2004) Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. Genetics 168: 435-446

Ladizinsky, G (1979) The origin of lentil and its wild genepool. Euphytica 28: 179-187.

Ladizinsky G, Cohen D, Muehlbauer F J (1985) Hybridization in the genus *Lens* by means of embryo culture. Theor Appl Genet 70: 97-101

Lázaro A, Ruiz M, de la Rosa L, Martín I (2001) Relationships between agro/morphological characters and climatic parameters in Spanish landraces of lentil (*Lens culinaris* Medik.). Gen Res Crop Evol 48: 239-249

Le BH, Wagmaister JA, Kawashima T, Bui AQ, Harada JJ, Goldberg RB (2007) Using genomics to study legume seed development. Pl Physiol 144: 562-574

Liu J, Guan JP, Xu DX, Zhang XY, Gu J, Zong XX (2008) Genetic diversity and population structure in lentil (*Lens culinaris* Medik.) germplasm detected by SSR markers. Acta Agron Sin 34: 1901-1909

Mackay I, Powell W (2007) Methods for linkage disequilibrium mapping in crops. Trends in Plant Science 12: 57-63

Manenti G, Galvan A, Pettinicchio A, Trincucci G, Spada E, Zolin A, Milani S, Gonzalez-Neira A, Dragani TA (2009) Mouse genome-wide association mapping needs linkage analysis to avoid false-positive loci. PLoS Genetics 5: e1000331

Morrall, RAA (1997) Evolution of lentil diseases over 25 years in western Canada. Can J Plant Path 19: 197-207

Muehlbauer FJ, Kaiser W J, Clement S L, Summerfield R J (1995) Production and breeding of lentil. Adv Agron 54: 283-332

Muñoz-Amatriaín M, Moscou MJ, Bhat PR, Svensson JT, Bartoš J, Suchánková P,Doležel J, Close TJ (2011) An improved consensus linkage map of barley based on flow-sorted chromosomes and single nucleotide polymorphism markers. The Plant Genome 3: 238-249

Nei, M (1972) Genetic distance between populations. Amer Naturalist 106: 283-292.

Neumann K, Kobiljski B, Denčić S, Varshney R K, Börner A (2011) Genome-wide association mapping: a case study in bread wheat (*Triticum aestivum* L.). Mol Breed 27: 37-58

Nordborg M, Borevitz J O, Bergelson J, Berry C C, Chory J, Hagenblad J, et al (2002) The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nature Genetics 30: 190-193

O'Boyle PD, Kelly JD, Kirk WW (2007) Use of marker-assisted selection to breed for resistance to common bacterial blight in common bean. J Amer Soc Hort Sci 132: 381-386

Ohto MA, Fischer RL, Goldberg RB, Nakamura K, Harada JJ (2005) Control of seed mass by APETALA2. Proc Nat Acad Sci Uni Stat Amer 102: 3123-3128

Oraguzie NC, Wilcox PL, Rikkerink E (2007) Linkage disequilibrium. In: Oraguzie NA, Rikkerink EHA, Gardiner SE, de Silva N (ed) Association mapping in plants. Springer, New York, pp 10-39

Page RD (1996) TreeView. An application to display phylogenetic trees on personal computer. Comp Appl Biol Sci 12: 357-358

Paran I, Goldman I, Tanksley SD, Zamir D (1995) Recombinant inbred lines for genetic mapping in tomato. Theor Appl Genet 90: 542-548

Pérez-Vega E, Pañeda A, Rodríguez-Suárez C, Campa A, Giraldez R, Ferreira JJ (2010) Mapping of QTLs for morpho-agronomic and seed quality traits in a RIL population of common bean (*Phaseolus vulgaris* L.). Theor Appl Genet 7: 1367-1380

Phan HTT, Ellwood SR, Hane JK, Ford R, Materne M, Oliver RP (2007) Extensive macrosynteny between *Medicago truncatula* and *Lens culinaris* ssp. *culinaris*. Theor Appl Genet 114: 549-558

Pinheiro JC, Bates DM (2000) Mixed-Effects Models in S and S-PLUS. Springer 27: 29

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38: 904-909

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945-959

Qiu X, Gong R, Tan Y, Yu S (2012) Mapping and characterization of the major quantitative trait locus *qSS7* associated with increased length and decreased width of rice seeds. Theor Appl Genet 125:1717-1726

R Development Core Team (2011) R: A language and environment for statistical computing. R foundation for statistical computing. http://R-project.org/. Accessed November, 2011

Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. Current Opinion in Plant Biology 5: 94-100

Robbins MD, Sim SC, Yang W, Van Deynze A, van der Knaap E, Joobeur T, Francis DM (2011) Mapping and linkage disequilibrium analysis with a genome-wide collection of SNPs that detect polymorphism in cultivated tomato. J Exper Bot 62: 1831-1845

Rodríguez GR, Moyseenko JB, Robbins MD, Morejón NH, Francis DM, van der Knaap, E. (2010) Tomato Analyzer: a useful software application to collect accurate and detailed morphological and colorimetric data from two-dimensional objects. J Visual Exper 37: 1856

Rubeena, Ford R, Taylor PWJ (2003) Construction of an intraspecific linkage map of lentil (*Lens culinaris* ssp. *culinaris*). Theor Appl Genet 107: 910-916

Saha G C, Sarker A, Chen W, Vandemark G J, Muehlbauer F J (2010) Identification of markers associated with genes for rust resistance in *Lens culinaris* Medik. Euphytica 175: 261-265

Salas P, Oyarzo-Llaipen JC, Wang D, Chase K, Mansur L (2006) Genetic mapping of seed shape in three populations of recombinant inbred lines of soybean (*Glycine max* L. Merr.). Theor Appl Genet 113: 1459-1466

Sandhu JS, Singh S (2007) History and origin. In: Yadav SS, McNeil DL, Stevenson, PC. (eds) Lentil: An ancient crop in modern times, Springer, Dordrecht, The Netherlands, pp 1-9

Sarker A, Erskine W, Sharma B, Tyagi MC (1999) Inheritance and linkage relationship of days to flower and morphological loci in lentil (*Lens culinaris* Medikus subsp. *culinaris*). J Hered 90: 270-275

Saskatchewan Pulse Growers (2011) http://www.saskpulse.com/. Accessed December, 2011

Semagn K, Bjornstad A, Xu Y (2010) The genetic dissection of quantitative traits in crops. Elec J Biotech doi: 10.2225/vol13-issue5-fulltext-21

Shahin MA, Symons SJ (2001) A machine vision system for grading lentils. Can Biosys Eng 43: 7-7

Shanin MA, Symons SJ, Poysa VW (2006) Determining soya bean size uniformity using image analysis. Biosys Eng 94: 191-198

Shahin MA, Symons SJ, Wang N (2012) Predicting dehulling efficiency of lentils based on seed size and shape characteristics measured with image analysis. Quality Assurance and Safety of Crops and Foods 4: 9-16

Sharma SK, Dawson IK, Waugh R (1995) Relationships among cultivated and wild lentils revealed by RAPD analysis. Theor Appl Genet 91: 647-654

Sharpe A, Ramsay L, Sanderson LA, Fedoruk M, Clarke W, Li R, Kagale S, Vijayan P, Vandenberg A, Bett K (2013) Ancient crop joins modern era: gene-based SNP discovery and mapping in lentil. BMC Genomics 14: 192

Singh SP (1994) Gamete selection for simultaneous improvement of multiple traits in common bean. Crop Sci 34: 352-355

Slinkard AE (1978) Inheritance of cotyledon color in lentils. J Hered 69: 139-140

Slinkard AE, Bhatty RS (1979) Laird lentil. Can J Plant Sci 59: 503-504

Slinkard AE (1981) Eston lentil. Can J Plant Sci 61: 733-734

Slinkard AE, Vandenebrg A (1995) Lentil. In: Slinkard A, Knott AE (ed) Harvest of gold: The history of field crop breeding in Canada. University Extension Press, Saskatoon, SK, pp 191-196

Somers DJS , Banks TB, DePauw RD, Fox SF, Clarke JC, Pozniak CP, McCartney CM (2007) Genome-wide linkage disequilibrium analysis in bread wheat and durum wheat. Genome 50: 557-567

Sonnante G, Pignone D (2001) Assessment of genetic variation in a collection of lentil using molecular tools. Euphytica 120: 301-307

Sonnante G, Hammer K, Pignone D (2009) From the cradle of agriculture a handful of lentils: history of domestication. Rendiconti Lincei 20: 21-37

Soto-Cerda BJ, Cloutier S (2012) Association mapping in plant genomes. In: Caliskan M (ed) Genetic diversity in plants. InTech, Rijeka, pp 29–54

Tadmor Y, Zamir D, Ladizinsky G (1987) Genetic mapping of an ancient translocation in the genus *Lens*. Theor Appl Genet 73:883–892

Tahir M, Simon CJ, Muehlbauer FJ (1993) Gene map of lentil: a review. *Lens* Newsletter 20: 3-10

Tahir M, Ghafoor A, Zubair M (1995) Genetics of seed weight in lentil (*Lens culinaris* Medik.). Pak J Bot 27: 435-440

Tahir M, Båga M, Vandenberg A, Chibbar RN (2012) An Assessment of Raffinose Family Oligosaccharides and Sucrose Concentration in Genus *Lens*. Crop Sci 52: 1713-1720

Tanabata T, Shibaya T, Hori K, Ebana K, Yano M (2012) SmartGrain: High-throughput phenotyping software for measuring seed shape through image analysis. Plant Physiology 160: 1871-1880

Tanyolac B, Ozatay S, Kahraman A, Muehlbauer F (2010) Linkage mapping of lentil (*Lens culinaris* L.) genome using recombinant inbred lines revealed by AFLP, ISSR, RAPD and some morphologic markers. J Agric Biotech Sustain Dev 2: 1-6

Tar'an B, Buchwaldt L, Tullu A, Banniza S, Warkentin TD, Vandenberg A (2003) Using molecular markers to pyramid genes for resistance to ascochyta blight and anthracnose in lentil (*Lens culinaris* Medik). Euphytica 134: 223-230

Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, Gaut BS (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Proc Nat Acad Sci 98: 9161-9166

Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES (2001) Dwarf8 polymorphisms associate with variation in flowering time. Nature Genetics 28: 286-289

Thoquet P, Ghérardi M, Journet EP, Kereszt A, Ané JM, Prosperi JM, Huguet T (2002) The molecular genetic linkage map of the model legume *Medicago truncatula*: an essential tool for comparative legume genomics and the isolation of agronomically important genes. BMC Plant Biology 2: 1

Tullu A, Kusmenoglu I, McPhee KE, Muehlbauer FJ (2001) Characterization of core collection of lentil germplasm for phenology, morphology, seed and straw yields. Gen Res Crop Evol 48: 143-152

Tullu A, Tar'an B, Breitkreutz C, Banniza S, Warkentin TD, Vandenberg A, Buchwaldt L (2006) A quantitative-trait locus for resistance to ascochyta blight (*Ascochyta lentis*) maps close to a gene for resistance to anthracnose (*Colletotrichum truncatum*) in lentil. Can J Pl Path 28: 588-595

Tullu A, Tar'an B, Warkentin T, Vandenburg A (2008) Construction of an intraspecific linkage map and QTL analysis for earliness and plant height in lentil. Crop Sci 48: 2254-2264

Tullu A, Banniza S, Tar'an B, Warkentin T, Vandenberg A (2010) Sources of resistance to ascochyta blight in wild species of lentil (*Lens culinaris* Medik.). Genet Res Crop Evol 57: 1053-1063

Vail S (2010) Interspecific-derived and juvenile resistance to anthracnose in lentil. Dissertation. University of Saskatchewan

Vandenberg A, Slinkard AE (1990) Genetics of seed coat color and pattern in lentil. J. Hered 81: 484-488

Vandenberg A, Kiehn FA, Vera C, Gaudiel R, Buchwaldt L, Dueck S, Wahab J, Slinkard AE (2002) CDC Robin lentil. Can J Plant Sci 82: 111-112

Van Berloo R (2008) GGT 2.0: Versatile software for visualization and analysis of genetic data. J Hered 99:232-236

van Ooijen JW, Voorips RE (2004) JoinMap Version 3.0, software for the calculation of genetic linkage maps. Kyazma BV, Wageningen, Netherlands

van Oss H, Arnon Y, Ladizinsky G (1997) Chloroplast DNA variation and evolution in the genus *Lens* Mill. Theor App Genet 94: 452-457

Wang N (2008) Effect of variety and crude protein content on dehulling quality and on the resulting chemical composition of red lentil (*Lens culinaris*). J Sci Food Ag 88: 885-890

Wang B, Jin SH, Hu HQ, Sun YG, Wang YW, Han P, Hou BK (2012) UGT87A2, an *Arabidopsis* glycosyltransferase, regulates flowering time via FLOWERING LOCUS C. New Phytologist 194: 666–675

Watanabe S, Tajuddin T, Yamanaka N, Hayashi M, Harada K (2004) Analysis of QTLs for reproductive development and seed quality traits in soybean using recombinant inbred lines. Breed Sci 4: 399-407

Weller JL, Hecht V, Liew LC, Sussmilch FC, Wenden B, Knowles CL, Vander Schoor, J K (2009) Update on the genetic control of flowering in garden pea. J Exp Bot 60: 2493-2499

Wiebe K, Harris NS, Faris JD, Clarke JM, Knox RE, Taylor GJ, Pozniak CJ (2010) Targeted mapping of Cdu1, a major locus regulating grain cadmium concentration in durum wheat (*Triticum turgidum* L. var *durum*). Theor Appl Genet 121: 1047-1058

Xu Y, Li HN, Li GJ, Wang X, Cheng LG, Zhang YM (2011) Mapping quantitative trait loci for seed size traits in soybean (*Glycine max* L. Merr.). Theor Appl Genet 122: 581-594

Xian-Liang S, Xue-Zhen S, Tian-Zhen Z (2006) Segregation distortion and its effect on genetic mapping in plants. Chinese J Agri Biotech 3: 163-169

Yan J, Yang X, Shah T, Sanchez-Villeda H, Li J, Warburton M, Zhou Y, Crouch JH, Xu Y (2009) High-throughput SNP genotyping with the GoldenGate assay in maize. Mol Breed 25: 441-451

Yu J, Pressoir G, Briggs W H, Bi I V, Yamasaki M, Doebley J F, et al (2005) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics 38: 203-208

Yu J, Buckler ES (2006) Genetic association mapping and genome organization of maize. Current Opinion in Biotechnology 17: 155-160

Zamir D, Ladizinsky G (1984) Genetics of allozyme variants and linkage groups in lentil. Euphytica 33: 329-336

Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, et al (2007) An *Arabidopsis* example of association mapping in structured samples. PLoS Genetics 3: e4

Appendices

Appendix 1. Association panel number, accession name, country of origin and their STRUCTURE sub-group assignment.

Number	Accession	Origin	Sub-group Assignment
1	2861-15a	Canada	1
2	3155-18	Canada	1
3	3156-11	Canada	1
4	3339-3	Canada	3
5	CDC Blaze	Canada	1
6	CDC Cherie	Canada	2
7	CDC Dazil	Canada	1
8	CDC Glamis	Canada	3
9	CDC Gold	Canada	3
10	CDC Grandora	Canada	3
11	CDC Greenland	Canada	3
12	CDC Imax	Canada	1
13	CDC Imigreen	Canada	4
14	CDC Impact	Canada	1
15	CDC Impala	Canada	1
16	CDC Imperial	Canada	1
17	CDC Impower	Canada	3
18	CDC Impress	Canada	3
19	CDC Improve	Canada	3
20	CDC Imvincible	Canada	1
21	CDC KR-1	Canada	3
22	CDC LeMay	Canada	1
23	CDC Matador	Canada	1
24	CDC Maxim	Canada	1
25	CDC Meteor	Canada	1
26	CDC Milestone	Canada	1
27	CDC Peridot	Canada	3
28	CDC Plato	Canada	4
29	CDC QG-1	Canada	1
30	CDC Red Rider	Canada	3
31	CDC Redberry	Canada	2
32	CDC Redbow	Canada	1
33	CDC Redcap	Canada	1
34	CDC Redcliff	Canada	1
35	CDC Redcoat	Canada	2
36	CDC Redwing	Canada	3
37	CDC Richlea	Canada	3
38	CDC Robin	Canada	1

39	CDC Rosebud	Canada	1
40	CDC Rosetown	Canada	1
41	CDC Rouleau	Canada	3
42	CDC Royale	Canada	1
43	CDC Ruby	Canada	1
44	CDC SB-1	Canada	1
45	CDC Sedley	Canada	3
46	CDC Sovereign	Canada	3
47	CDC Vantage	Canada	3
48	CDC Viceroy	Canada	1
49	Crimson	USA	1
50	Eston	Canada	1
51	ILL 0009	Jordan	2
52	ILL 0028	Syria	2
53	ILL 0080	Spain	2
54	ILL 0242	Iran	4
55	ILL 0293	Greece	2
56	ILL 0313	Palestine	2
57	ILL 0618	Tajikistan	4
58	ILL 0624	Macedonia	1
59	ILL 0927	Turkey	4
60	ILL 1139	Lebanon	2
61	ILL 1220	Iran	4
62	ILL 1337	Iran	4
63	ILL 1553	Iran	4
64	ILL 1762	Afghan	4
65	ILL 1861	Sudan	4
66	ILL 1983	Ethiopia	1
67	ILL 2194	Russia	3
68	ILL 2217	Afghan	4
69	ILL 2290	Chile	2
70	ILL 2433	Ethiopia	2
71	ILL 2501	India	4
72	ILL 2526	India	4
73	ILL 2607	India	4
74	ILL 2684	India	4
75	ILL 2789	India	4
76	ILL 3025	India	4
77	ILL 3347	India	4
78	ILL 3502	Ukraine	1
79	ILL 3597	India	4
80	ILL 4164	India	4
81	ILL 4359	India	4

82	ILL 4400	Syria	2
83	ILL 4605	Argentina	2
84	ILL 4609	Netherlands	4
85	ILL 4665	Hungary	1
86	ILL 4671	USA	2
87	ILL 4740	France	1
88	ILL 4768	Yemen	2
89	ILL 4783	Czech Republic	2
90	ILL 4804	Libya	1
91	ILL 4875	Uzebekistan	4
92	ILL 4956	Portugal	1
93	ILL 5058	Spain	2
94	ILL 5151	India	2
95	ILL 5209	Jordan	2
96	ILL 5511	Syria	2
97	ILL 5576	Serbia	2
98	ILL 5588	Jordan	2
99	ILL 5883	Jordan	2
100	ILL 5945	Ethiopia	1
101	ILL 6182	Tunisia	1
102	ILL 6853	Syria	2
103	ILL 6967	Brazil	4
104	ILL 7051	Algeria	2
105	ILL 7089	Russia	1
106	ILL 7585	Turkey	2
107	ILL 7747	Syria	1
108	Indian Head	Canada	1
109	Laird	Canada	3
110	PI 178939	Turkey	1
111	PI 178971	Turkey	4
112	PI 217949	Pakistan	4
113	PI 251032	Iran	4
114	PI 273664	Ethiopia	2
115	PI 297284	Argentina	1
116	PI 298631	Peru	2
117	PI 298922	Italy	1
118	PI 299121	Mexico	1
119	PI 299126	Mexico	1
120	PI 299215	Chile	4
121	PI 300250	Syria	4
122	PI 308614	Syria	4
123	PI 320954	Hungary	1
124	PI 329169	Iran	2

125	PI 339283	Turkey	4
126	PI 339285	Turkey	4
127	PI 339292	Turkey	1
128	PI 343026	Former Soviet Union	4
129	PI 357225	Former Serbia and Montenegro	1
130	PI 368647	Macedonia	1
131	PI 426803	Pakistan	4
132	PI 431662	Iran	4
133	PI 431679	Iran	4
134	PI 431705	Iran	4
135	PI 431710	Iran	4
136	PI 431714	Iran	4
137	PI 431717	Iran	4
138	PI 431756	Iran	4
139	W6 27764	USA	4
140	W6 27766	USA	4

Appendix 2. STRUCTURE sub-groups and each line's assigned groupings. Each value within the sub-groups shows the level of admixture that each accession has for the sub-groups.

		STRUCTURE sub-groups				Sub-Group
Number	Accession	1	2	3	4	Assignment
1	2861-15a	0.934	0	0.065	0.001	1
2	3155-18	0.918	0.081	0	0	1
3	3156-11	0.478	0.136	0.385	0.001	1
4	3339-3	0.03	0.042	0.927	0.001	3
5	CDC CDC Blaze	0.78	0.119	0.101	0.001	1
6	CDC CDC Cherie	0.001	0.897	0	0.102	2
7	CDC Dazil	0.735	0.076	0.186	0.002	1
8	CDC Glamis	0	0	1	0	3
9	CDC Gold	0.109	0.001	0.75	0.14	3
10	CDC Grandora	0	0	1	0	3
11	CDC Greenland	0.001	0.056	0.942	0.001	3
12	CDC Imax	0.998	0	0.001	0	1
13	CDC Imigreen	0	0.269	0.002	0.728	4
14	CDC Impact	0.792	0.113	0.094	0.001	1
15	CDC Impala	0.986	0.005	0.009	0	1
16	CDC Imperial	0.999	0	0	0.001	1
17	CDC Impower	0.189	0.044	0.766	0.001	3
18	CDC Impress	0.021	0.057	0.92	0.002	3
19	CDC Improve	0.008	0.107	0.885	0.001	3
20	CDC Imvincible	0.999	0	0.001	0	1
21	CDC KR-1	0.149	0.219	0.632	0	3
22	CDC LeMay	0.995	0	0.004	0	1
23	CDC Matador	0.965	0.032	0.002	0.001	1
24	CDC Maxim	0.447	0.394	0.159	0	1
25	CDC Meteor	0.996	0	0.001	0.002	1
26	CDC Milestone	0.646	0.03	0.323	0	1
27	CDC Peridot	0	0	0.999	0	3
28	CDC Plato	0.024	0	0.001	0.975	4
29	CDC QG-1	0.563	0	0.436	0.001	1
30	CDC Red_Rider	0.071	0.199	0.707	0.023	3
31	CDC Redberry	0.357	0.387	0.256	0	2
32	CDC Redbow	0.905	0.08	0.014	0.001	1
33	CDC Redcap	0.564	0.151	0.285	0.001	1
34	CDC Redcliff	0.547	0.25	0.203	0	1
35	CDC Redcoat	0.331	0.366	0.302	0.001	2
36	CDC Redwing	0	0.424	0.575	0	3
37	CDC Richlea	0.358	0.001	0.641	0	3
38	CDC Robin	0.999	0	0	0	1

39	CDC Rosebud	0.998	0	0	0.001	1
40	CDC Rosetown	0.916	0.083	0.001	0	1
41	CDC Rouleau	0.388	0	0.61	0.002	3
42	CDC Royale	0.698	0	0.301	0.001	1
43	CDC Ruby	0.866	0	0.133	0	1
44	CDC SB-1	0.73	0.163	0.002	0.105	1
45	CDC Sedley	0	0.168	0.831	0	3
46	CDC Sovereign	0.129	0.012	0.857	0.002	3
47	CDC Vantage	0.076	0.008	0.884	0.033	3
48	CDC Viceroy	0.9	0.041	0.008	0.051	1
49	Crimson	0.749	0.134	0.036	0.08	1
50	Eston	0.888	0.086	0.025	0.001	1
51	ILL 0009	0.001	0.982	0.015	0.002	2
52	ILL 0028	0.003	0.986	0.008	0.003	2
53	ILL 0080	0.127	0.795	0.078	0	2
54	ILL 0242	0.113	0.009	0.001	0.877	4
55	ILL 0293	0.109	0.869	0.001	0.02	2
56	ILL 0313	0.203	0.58	0.092	0.125	2
57	ILL 0618	0.049	0.03	0.003	0.918	4
58	ILL 0624	0.999	0	0	0	1
59	ILL 0927	0.165	0.144	0.004	0.686	4
60	ILL 1139	0.002	0.752	0.001	0.246	2
61	ILL 1220	0.235	0.238	0.006	0.52	4
62	ILL 1337	0.027	0.057	0.004	0.912	4
63	ILL 1553	0.313	0.229	0.003	0.455	4
64	ILL 1762	0.001	0.001	0.001	0.998	4
65	ILL 1861	0.138	0.162	0	0.7	4
66	ILL 1983	0.892	0.076	0.001	0.031	1
67	ILL 2194	0	0	0.999	0	3
68	ILL 2217	0.237	0.063	0.065	0.634	4
69	ILL 2290	0.349	0.42	0.229	0.001	2
70	ILL 2433	0	0.999	0	0.001	2
71	ILL 2501	0	0	0	0.999	4
72	ILL 2526	0	0	0	1	4
73	ILL 2607	0	0	0	0.999	4
74	ILL 2684	0	0	0	0.999	4
75	ILL 2789	0	0	0	0.999	4
76	ILL 3025	0	0	0	0.999	4
77	ILL 3347	0	0	0	0.999	4
78	ILL 3502	0.447	0.378	0.003	0.172	1
79	ILL 3597	0	0	0	0.999	4
80	ILL 4164	0.173	0.213	0.001	0.612	4
81	ILL 4359	0	0	0	0.999	4

82	ILL 4400	0.001	0.997	0.001	0	2
83	ILL 4605	0.349	0.507	0.143	0.001	2
84	ILL 4609	0	0	0	0.999	4
85	ILL 4665	0.338	0.337	0.322	0.004	1
86	ILL 4671	0.087	0.785	0.128	0.001	2
87	ILL 4740	0.999	0	0.001	0	1
88	ILL 4768	0	0.999	0	0.001	2
89	ILL 4783	0	0.999	0	0.001	2
90	ILL 4804	0.509	0.32	0.002	0.169	1
91	ILL 4875	0.004	0.028	0.021	0.946	4
92	ILL 4956	0.863	0.114	0.021	0.003	1
93	ILL 5058	0.204	0.693	0.103	0	2
94	ILL 5151	0	0.951	0	0.048	2
95	ILL 5209	0.002	0.987	0.004	0.007	2
96	ILL 5511	0.152	0.734	0.114	0	2
97	ILL 5576	0.297	0.677	0.02	0.006	2
98	ILL 5588	0	0.999	0	0.001	2
99	ILL 5883	0	0.999	0	0	2
100	ILL 5945	0.551	0.219	0.223	0.007	1
101	ILL 6182	0.555	0.239	0.006	0.2	1
102	ILL 6853	0.027	0.96	0.002	0.011	2
103	ILL 6967	0.12	0.411	0.003	0.467	4
104	ILL 7051	0.237	0.432	0.094	0.236	2
105	ILL 7089	0.544	0.097	0.35	0.009	1
106	ILL 7585	0.164	0.482	0.002	0.352	2
107	ILL 7747	0.955	0	0.042	0.002	1
108	Indian Head	0.999	0	0	0	1
109	Laird	0	0	0.999	0	3
110	PI 178939	0.942	0	0	0.058	1
111	PI 178971	0.001	0	0.001	0.998	4
112	PI 217949	0	0	0	0.999	4
113	PI 251032	0.252	0.079	0.003	0.666	4
114	PI 273664	0	0.999	0	0.001	2
115	PI 297284	0.512	0.361	0.126	0.001	1
116	PI 298631	0.012	0.774	0.152	0.061	2
117	PI 298922	0.595	0.226	0.011	0.167	1
118	PI 299121	0.426	0.355	0.169	0.05	1
119	PI 299126	0.532	0.197	0.164	0.108	1
120	PI 299215	0.439	0.085	0.01	0.467	4
121	PI 300250	0.139	0.22	0.001	0.641	4
122	PI 308614	0.144	0.209	0.001	0.646	4
123	PI 320954	0.964	0.001	0.034	0.001	1
124	PI 329169	0.116	0.548	0.022	0.313	2

125	PI 339283	0.151	0.156	0.001	0.692	4
126	PI 339285	0.16	0.122	0.001	0.717	4
127	PI 339292	0.535	0.073	0.272	0.12	1
128	PI 343026	0.403	0.093	0.021	0.483	4
129	PI 357225	0.939	0.008	0.052	0.002	1
130	PI 368647	0.996	0	0.001	0.004	1
131	PI 426803	0.015	0.002	0	0.983	4
132	PI 431662	0	0	0	0.999	4
133	PI 431679	0.001	0	0	0.998	4
134	PI 431705	0	0	0	0.999	4
135	PI 431710	0.001	0	0	0.998	4
136	PI 431714	0.001	0.001	0	0.998	4
137	PI 431717	0	0.001	0	0.999	4
138	PI 431756	0.001	0	0	0.998	4
139	W6 27764	0	0	0	1	4
140	W6 27766	0.238	0.298	0.041	0.423	4